

Sparse Methods for Direction-of-Arrival Estimation

Zai Yang^{*†}, Jian Li[‡], Petre Stoica[§], and Lihua Xie[†]

October 3, 2016

Contents

1	Introduction	3
2	Data Model	4
2.1	Data Model	4
2.2	The Role of Array Geometry	4
2.3	Parameter Identifiability	6
3	Sparse Representation and DOA estimation	7
3.1	Sparse Representation and Compressed Sensing	7
3.1.1	Problem Formulation	7
3.1.2	Convex Relaxation	8
3.1.3	ℓ_q Optimization	10
3.1.4	Maximum Likelihood Estimation (MLE)	11
3.2	Sparse Representation and DOA Estimation: the Link and the Gap	12
4	On-Grid Sparse Methods	13
4.1	Data Model	13
4.2	$\ell_{2,0}$ Optimization	14
4.3	Convex Relaxation	14
4.3.1	$\ell_{2,1}$ Optimization	14
4.3.2	Dimensionality Reduction via $\ell_{2,1}$ -SVD	15
4.3.3	Another Dimensionality Reduction Technique	16
4.4	$\ell_{2,q}$ Optimization	17
4.5	Sparse Iterative Covariance-based Estimation (SPICE)	17
4.5.1	Generalized Least Squares	17
4.5.2	SPICE	18
4.6	Maximum Likelihood Estimation	21
4.7	Remarks on Grid Selection	21

^{*}School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China

[†]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

[‡]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA

[§]Department of Information Technology, Uppsala University, Uppsala, SE 75105, Sweden

5	Off-Grid Sparse Methods	22
5.1	Fixed Grid	22
5.1.1	Data Model	22
5.1.2	ℓ_1 Optimization	23
5.1.3	Sparse Bayesian Learning	24
5.2	Dynamic Grid	24
5.2.1	Data Model	24
5.2.2	Algorithms	24
6	Gridless Sparse Methods	25
6.1	Data Model	26
6.2	Vandermonde Decomposition of Toeplitz Covariance Matrices	26
6.3	The Single Snapshot Case	29
6.3.1	A General Framework for Deterministic Methods	29
6.3.2	Atomic ℓ_0 Norm	29
6.3.3	Atomic Norm	30
6.3.4	Hankel-based Nuclear Norm	33
6.3.5	Connection between ANM and EMaC	34
6.3.6	Covariance Fitting Method: Gridless SPICE (GLS)	35
6.3.7	Connection between ANM and GLS	37
6.4	The Multiple Snapshot Case: Covariance Fitting Methods	37
6.4.1	Gridless SPICE (GLS)	38
6.4.2	SMV-based Atomic Norm Minimization (ANM-SMV)	40
6.4.3	Nuclear Norm Minimization followed by MUSIC (NNM-MUSIC)	40
6.4.4	Comparison of GLS, ANM-SMV and NNM-MUSIC	41
6.5	The Multiple Snapshot Case: Deterministic Methods	41
6.5.1	A General Framework	41
6.5.2	Atomic ℓ_0 Norm	42
6.5.3	Atomic Norm	43
6.5.4	Hankel-based Nuclear Norm	45
6.6	Reweighted Atomic Norm Minimization	46
6.6.1	A Smooth Surrogate for $\ \mathbf{Z}\ _{\mathcal{A},0}$	46
6.6.2	A Locally Convergent Iterative Algorithm	47
6.6.3	Interpretation as RAM	48
6.7	Connections between ANM and GLS	49
6.7.1	The Case of $L < M$	49
6.7.2	The Case of $L \geq M$	50
6.8	Computational Issues and Solutions	51
6.8.1	Dimensionality Reduction	51
6.8.2	Alternating Direction Method of Multipliers (ADMM)	52
7	Future Research Challenges	53
8	Conclusions	54

1 Introduction

Direction of arrival (DOA) estimation refers to the process of retrieving the direction information of several electromagnetic waves/sources from the outputs of a number of receiving antennas that form a sensor array. DOA estimation is a major problem in array signal processing and has wide applications in radar, sonar, wireless communications, etc.

The study of DOA estimation methods has a long history. For example, the conventional (Bartlett) beamformer, which dates back to the World War II, simply uses Fourier-based spectral analysis of the spatially sampled data. Capon's beamformer was later proposed to improve the estimation performance of closely spaced sources [1]. Since the 1970s when Pisarenko found that the DOAs can be retrieved from data second order statistics [2], a prominent class of methods designated as subspace-based methods have been developed, e.g., the multiple signal classification (MUSIC) and the estimation of parameters by rotational invariant techniques (ESPRIT) along with their variants [3–7]. Another common approach is the nonlinear least squares (NLS) method that is also known as the (deterministic) maximum likelihood estimation. For a complete review of these methods, readers are referred to [8,9]. Note that these methods suffer from certain well-known limitations. For example, the subspace-based methods and the NLS need *a priori* knowledge on the source number that may be difficult to obtain. Additionally, Capon's beamformer, MUSIC and ESPRIT are covariance-based and require a sufficient number of data snapshots to accurately estimate the data covariance matrix. Moreover, they can be sensitive to source correlations that tend to cause a rank deficiency in the sample data covariance matrix. Also, a very accurate initialization is required for the NLS since its objective function has a complicated multimodal shape with a sharp global minimum.

The purpose of this article is to provide an overview of the recent work on sparse DOA estimation methods. These new methods are motivated by techniques in sparse representation and compressed sensing methodology [10–14], and most of them have been proposed during the last decade. The sparse estimation (or optimization) methods can be applied in several demanding scenarios, including cases with no knowledge of the source number, limited number of snapshots (even a single snapshot), and highly or completely correlated sources. Due to these attractive properties they have been extensively studied and their popularity is reflected by the large number of publications about them.

It is important to note that there is a key difference between the common sparse representation framework and DOA estimation. To be specific, the studies of sparse representation have been focused on *discrete linear* systems. In contrast to this, the DOA parameters are *continuous* valued and the observed data are *nonlinear* in the DOAs. Depending on the model adopted, we can classify the sparse methods for DOA estimation into three categories, namely, *on-grid*, *off-grid* and *gridless*, which also corresponds to the chronological order in which they have been developed. For on-grid sparse methods, the data model is obtained by assuming that the true DOAs lie on a set of fixed grid points in order to straightforwardly apply the existing sparse representation techniques. While a grid is still required by off-grid sparse methods, the DOAs are not restricted to be on the grid. Finally, the recent gridless sparse methods do not need a grid, as their name suggests, and they operate directly in the continuous domain.

The organization of this article is as follows. The data model for DOA estimation is introduced in Section 2 for far-field, narrowband sources that are the focus of this article. Its dependence on the array geometry and the parameter identifiability problem are discussed. In Section 3 the concepts of sparse representation and compressed sensing are introduced and several sparse estimation techniques are discussed. Moreover, we discuss the feasibility of using sparse representation techniques for DOA estimation and highlight the key differences between sparse representation and DOA estimation. The on-grid sparse methods for DOA estimation are introduced in Section 4. Since they are straightforward to obtain in the case of a single data snapshot, we focus on showing how the temporal redundancy of multiple snapshots can be utilized to improve the DOA estimation performance. Then, the off-grid sparse methods are presented in Section 5. Section 6 is the main highlight of this article in which the recently developed gridless sparse methods are

presented. These methods are of great interest since they operate directly in the continuous domain and have strong theoretical guarantees. Some future research directions will be discussed in Section 7 and conclusions will be drawn in Section 8.

Notations used in this article are as follows. \mathbb{R} and \mathbb{C} denote the sets of real and complex numbers respectively. Boldface letters are reserved for vectors and matrices. $|\cdot|$ denotes the amplitude of a scalar or the cardinality of a set. $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the ℓ_1 , ℓ_2 and Frobenius norms respectively. \mathbf{A}^T , \mathbf{A}^* and \mathbf{A}^H are the matrix transpose, complex conjugate and conjugate transpose of \mathbf{A} respectively. x_j is the j th entry of a vector \mathbf{x} , and \mathbf{A}_j denotes the j th row of a matrix \mathbf{A} . Unless otherwise stated, \mathbf{x}_Ω and \mathbf{A}_Ω are the subvector and submatrix of \mathbf{x} and \mathbf{A} obtained by retaining the entries of \mathbf{x} and the rows of \mathbf{A} indexed by the set Ω . For a vector \mathbf{x} , $\text{diag}(\mathbf{x})$ is a diagonal matrix with \mathbf{x} on the diagonal. $\mathbf{x} \succeq \mathbf{0}$ means $x_j \geq 0$ for all j . $\text{rank}(\mathbf{A})$ denotes the rank of a matrix \mathbf{A} and $\text{Tr}(\mathbf{A})$ denotes the trace. For positive semidefinite matrices \mathbf{A} and \mathbf{B} , $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite. Finally, $\mathbb{E}[\cdot]$ denotes the expectation of a random variable, and for notational simplicity a random variable and its numerical value will not be distinguished.

2 Data Model

2.1 Data Model

In this section, the DOA estimation problem is stated. Consider K narrowband far-field source signals s_k , $k = 1, \dots, K$, impinging on an array of omnidirectional sensors from directions θ_k , $k = 1, \dots, K$. According to [8, 9], the time delays at different sensors can be represented by simple phase shifts, resulting in the following data model:

$$\mathbf{y}(t) = \sum_{k=1}^K \mathbf{a}(\theta_k) s_k(t) + \mathbf{e}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{e}(t), \quad t = 1, \dots, L, \quad (1)$$

where t indexes the snapshot and L is the number of snapshots, $\mathbf{y}(t) \in \mathbb{C}^M$, $\mathbf{s}(t) \in \mathbb{C}^K$ and $\mathbf{e}(t) \in \mathbb{C}^M$ denote the array output, the vector of source signals and the vector of measurement noise at snapshot t , respectively, where M is the number of sensors. $\mathbf{a}(\theta_k)$ is the so-called steering vector of the k th source that is determined by the geometry of the sensor array and will be given later. The steering vectors compose the array manifold matrix $\mathbf{A}(\boldsymbol{\theta}) := [\mathbf{a}(\theta_1), \dots, \mathbf{a}(\theta_K)]$. More compactly, (1) can be written as

$$\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{S} + \mathbf{E}, \quad (2)$$

where $\mathbf{Y} = [\mathbf{y}(1), \dots, \mathbf{y}(L)]$, and \mathbf{S} and \mathbf{E} are similarly defined. Given the data matrix \mathbf{Y} and the mapping $\theta \rightarrow \mathbf{a}(\theta)$, the objective is to estimate the parameters θ_k , $k = 1, \dots, K$ that are referred to as the DOAs. It is worth noting that the source number K is usually unknown in practice; typically, K is assumed to be smaller than M , as otherwise the DOAs cannot be uniquely identified from the data (see details in Subsection 2.3).

2.2 The Role of Array Geometry

We now discuss how the mapping $\theta \rightarrow \mathbf{a}(\theta)$ is determined by the array geometry. We first consider a general 2-D array with the M sensors located at points $(r_m, \check{\theta}_m)$, expressed in polar coordinates. For convenience, the unit of distance is taken as half the wavelength of the waves. Then $\mathbf{a}(\theta)$ will be given by

$$a_m(\theta) = e^{i\pi r_m \cos(\theta - \check{\theta}_m)}, \quad m = 1, \dots, M. \quad (3)$$

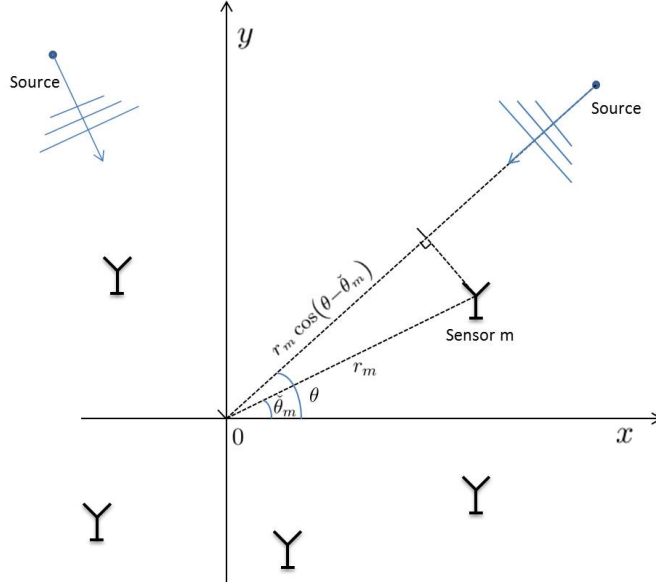


Figure 1: The setup of the DOA estimation problem with a general 2-D array.

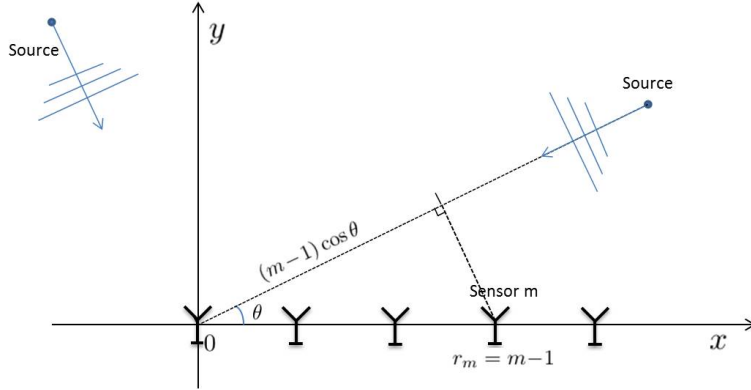


Figure 2: The setup of the DOA estimation problem with a ULA.

In the particularly interesting case of a linear array, assuming that the sensors are located on the nonnegative x -axis, we have that $\tilde{\theta}_m = 0^\circ$, $m = 1, \dots, M$. Therefore, $\mathbf{a}(\theta)$ will be given by

$$a_m(\theta) = e^{i\pi r_m \cos \theta}, \quad m = 1, \dots, M. \quad (4)$$

We can replace the variable θ by $f = \frac{1}{2} \cos \theta$ and define without ambiguity $\mathbf{a}(f) := \mathbf{a}(\arccos(2f)) = \mathbf{a}(\theta)$. Then, the mapping $\mathbf{a}(f)$ is given by

$$a_m(f) = e^{i2\pi r_m f}, \quad m = 1, \dots, M. \quad (5)$$

In the case of a single snapshot, obviously, the *spatial* DOA estimation problem becomes a *temporal* frequency estimation problem (a.k.a. line spectral estimation) given the samples y_m , $m = 1, \dots, M$ measured at time instants r_m , $m = 1, \dots, M$.

If we further assume that the sensors of the linear array are equally spaced, then the array is known as a uniform linear array (ULA). We consider the case when two adjacent antennas of the array are spaced by a

unit distance (half the wavelength). Then, we have that $r_m = m - 1$ and

$$\mathbf{a}(f) = \left[1, e^{i2\pi f}, \dots, e^{i2\pi(M-1)f} \right]^T. \quad (6)$$

If a linear array is obtained from a ULA by retaining only a part of the sensors, then it is known as a sparse linear array (SLA).

It is worth noting that for a 2-D array, it is possible to estimate the DOAs in the entire 360° range, while for a linear array we can only resolve the DOAs in a 180° range: $\theta_k \in [0^\circ, 180^\circ)$. In the latter case, correspondingly, the “frequencies” are: $f_k = \frac{1}{2} \cos \theta_k \in (-\frac{1}{2}, \frac{1}{2}]$. Throughout this article, we let \mathcal{D}_θ denote the domain of the DOAs that can be $[0^\circ, 360^\circ)$ or $[0^\circ, 180^\circ)$, depending on the context. Also, let $\mathbb{T} = (-\frac{1}{2}, \frac{1}{2}]$ be the frequency interval for linear arrays. Finally, we close this section by noting that the grid-based (on-grid and off-grid) sparse methods can be applied to arbitrary sensor arrays, while the existing gridless sparse methods are typically limited to ULAs or SLAs.

2.3 Parameter Identifiability

The DOAs $\{\theta_k\}_{k=1}^K$ can be uniquely identified from $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}) \mathbf{S}$ if there do not exist $\{\theta'_k\}_{k=1}^K \neq \{\theta_k\}_{k=1}^K$ and \mathbf{S}' such that $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}) \mathbf{S} = \mathbf{A}(\boldsymbol{\theta}') \mathbf{S}'$. Guaranteeing that the parameters can be uniquely identified in the noiseless case is usually a prerequisite for their accurate estimation. The parameter identifiability problem for DOA estimation was studied in [15] for ULAs and in [16, 17] for general arrays. The results in [18–20] are also closely related to this problem. For a general array, define the set

$$\mathcal{A}_\theta := \{\mathbf{a}(\theta) : \theta \in \mathcal{D}_\theta\}, \quad (7)$$

and define the spark of \mathcal{A}_θ , denoted by $\text{spark}(\mathcal{A}_\theta)$, as the smallest number of elements in \mathcal{A}_θ that are linearly dependent [21]. For any M -element array, it holds that

$$\text{spark}(\mathcal{A}_\theta) \leq M + 1. \quad (8)$$

Note that it is generally difficult to compute $\text{spark}(\mathcal{A}_\theta)$, except in the ULA case in which $\text{spark}(\mathcal{A}_\theta) = M + 1$ by the fact that any M steering vectors in \mathcal{A}_θ are linearly independent.

The paper [16] showed that any K sources can be uniquely identified from $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}) \mathbf{S}$ provided that

$$K < \frac{\text{spark}(\mathcal{A}_\theta) - 1 + \text{rank}(\mathbf{S})}{2}. \quad (9)$$

Note that the above condition cannot be easily used in practice since it requires knowledge on \mathbf{S} . To resolve this problem, it was shown in [19] that the condition in (9) is equivalent to

$$K < \frac{\text{spark}(\mathcal{A}_\theta) - 1 + \text{rank}(\mathbf{Y})}{2}. \quad (10)$$

Moreover, the condition in (9) or (10) is also necessary [19]. Combining these results, we have the following theorem.

Theorem 2.1. *Any K sources can be uniquely identified from $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}) \mathbf{S}$ if and only if the condition in (10) holds.*

Theorem 2.1 provides a necessary and sufficient condition for unique identifiability of the parameters. In the single snapshot case, the condition in (10) reduces to

$$K < \frac{\text{spark}(\mathcal{A}_\theta)}{2}. \quad (11)$$

Therefore, Theorem 2.1 implies that more sources can be determined if more snapshots are collected, except in the trivial case when the source signals are identical up to scaling factors. In the ULA case, the condition in (10) can be simplified as

$$K < \frac{M + \text{rank}(\mathbf{Y})}{2}. \quad (12)$$

Using the inequality $\text{rank}(\mathbf{S}) \leq K$ and (8), the condition in (9) or (10) implies that

$$K < \text{spark}(\mathcal{A}_\theta) - 1 \leq M. \quad (13)$$

Theorem 2.1 specifies the condition required to guarantee unique identifiability for *any* K source signals. It was shown in [16] that if $\{\theta_k\}$ are fixed and \mathbf{S} is randomly drawn from some absolutely continuous distribution, then the K sources can be uniquely identified with probability one, provided that

$$K < \frac{2\text{rank}(\mathbf{S})}{2\text{rank}(\mathbf{S}) + 1} (\text{spark}(\mathcal{A}_\theta) - 1). \quad (14)$$

Moreover, the following condition, which is slightly different from that in (14), is necessary:

$$K \leq \frac{2\text{rank}(\mathbf{S})}{2\text{rank}(\mathbf{S}) + 1} (\text{spark}(\mathcal{A}_\theta) - 1). \quad (15)$$

The condition in (14) is weaker than that in (9) or (10). As an example, in the single snapshot case, the upper bounds on K in (10) and (14) are approximately $\frac{1}{2}\text{spark}(\mathcal{A}_\theta)$ and $\frac{2}{3}\text{spark}(\mathcal{A}_\theta)$, respectively. However, the paper [17] pointed out that the condition in (14) has a relatively limited practical relevance in finite-SNR applications, since under (14), with a strictly positive probability, false DOA estimates far from the true DOAs may occur.

3 Sparse Representation and DOA estimation

In this section we will introduce the basics of sparse representation that has been an active research topic especially in the past decade. More importantly, we will discuss its connections to and the key differences from DOA estimation.

3.1 Sparse Representation and Compressed Sensing

3.1.1 Problem Formulation

We first introduce the topic of sparse representation and the closely related concept of compressed sensing that have found broad applications in image, audio and signal processing, communications, medical imaging, and computational biology, to name just a few (see, e.g., the various special issues in several journals [22–25]). Let $\mathbf{y} \in \mathbb{C}^M$ be the signal that we observe. We want to sparsely represent \mathbf{y} via the following model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \quad (16)$$

where $\mathbf{A} \in \mathbb{C}^{M \times \overline{N}}$ is a *given* matrix, with $M \ll \overline{N}$, that is referred as a dictionary and whose columns are called atoms, $\mathbf{x} \in \mathbb{C}^{\overline{N}}$ is a sparse coefficient vector (note that the notation N is reserved for later use), and \mathbf{e} accounts for the representation error. By sparsity we mean that only a few entries, say $K \ll \overline{N}$, of \mathbf{x} are nonzero and the rest are zero. This together with (16) implies that \mathbf{y} can be well approximated by a linear combination of K atoms in \mathbf{A} . The underlying motivation for the sparse representation is that even though the observed data \mathbf{y} lies in a high-dimensional space, it can actually be well approximated in some

lower-dimensional subspace ($K < M$). Given \mathbf{y} and \mathbf{A} , the problem of sparse representation is to find the sparse vector \mathbf{x} subject to data consistency.

The concept of sparse representation was later extended within the framework of compressed sensing [12–14]. In compressed sensing, \mathbf{x} is the sparse signal of interest that is acquired via the underdetermined linear system in (16). Given $M \ll \overline{N}$, the acquired signal \mathbf{y} is referred to as the compressive data, \mathbf{A} is the *given* sensing matrix, and \mathbf{e} denotes the measurement noise. Given \mathbf{y} and \mathbf{A} , the problem of sparse signal recovery in compressed sensing is also to solve for the sparse vector \mathbf{x} subject to data consistency. With no rise for ambiguity we will not distinguish between the terminologies used for sparse representation and compressed sensing, as these two problems are very much alike.

To solve for the sparse signal, intuitively, we should find the sparsest solution. In the absence of noise, therefore, we should solve the following optimization problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}, \quad (17)$$

where $\|\mathbf{x}\|_0 := \#\{j : x_j \neq 0\}$ counts the nonzero entries of \mathbf{x} and is referred to as the ℓ_0 (pseudo-)norm or the sparsity of \mathbf{x} . View \mathbf{A} as the set of its column vectors, and define its spark, denoted by $\text{spark}(\mathbf{A})$ as in Subsection 2.3. It can be shown that the true sparse signal \mathbf{x} can be uniquely determined by (17) if \mathbf{x} has a sparsity of

$$K < \frac{\text{spark}(\mathbf{A})}{2}. \quad (18)$$

To see this, suppose there exists \mathbf{x}' of sparsity $K' \leq K$ satisfying $\mathbf{y} = \mathbf{A}\mathbf{x}'$ as well. Then it holds that $\mathbf{A}(\mathbf{x} - \mathbf{x}') = \mathbf{0}$. Since $\mathbf{x} - \mathbf{x}'$ has a sparsity of at most $K + K' \leq 2K < \text{spark}(\mathbf{A})$, which holds following (18), it can be concluded that $\mathbf{x} - \mathbf{x}' = \mathbf{0}$ and thus $\mathbf{x} = \mathbf{x}'$ since any $\text{spark}(\mathbf{A}) - 1$ columns of \mathbf{A} are linearly independent. It is interesting to note that the condition in (18) is very similar to that in (11) required to guarantee identifiability for DOA estimation in the single snapshot case.

Unfortunately, the ℓ_0 optimization problem in (17) is NP hard to solve. Therefore, more efficient approaches are needed. We note that many methods and algorithms have been proposed for sparse signal recovery, e.g., convex relaxation or ℓ_1 optimization [10, 11], ℓ_q , $0 < q < 1$ (pseudo-)norm optimization [26–32], greedy algorithms such as orthogonal matching pursuit (OMP), compressive sampling matching pursuit (CoSaMP) and subspace pursuit (SP) [33–38], iterative hard thresholding (IHT) [39], maximum likelihood estimation (MLE), etc. Readers can consult [40] for a review. Here we introduce convex relaxation, ℓ_q optimization and MLE in the ensuing subsections.

3.1.2 Convex Relaxation

The first practical approach to sparse signal recovery that we will introduce is based on the convex relaxation, which replaces the ℓ_0 norm by its tightest convex relaxation—the ℓ_1 norm. Therefore, we solve the following optimization problem in lieu of (17):

$$\min \|\mathbf{x}\|_1, \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}, \quad (19)$$

which is sometimes referred to as basis pursuit (BP) [10]. Since the ℓ_1 norm is convex, (19) can be solved in a polynomial time. In fact, the use of ℓ_1 optimization for obtaining a sparse solution dates back to the paper [41] about seismic data recovery. While the BP was empirically observed to give good performance, a rigorous analysis had not been provided for decades.

To introduce the existing theoretical guarantees for the BP in (19), we define a metric of the matrix \mathbf{A} called mutual coherence that quantifies the correlations between the atoms in \mathbf{A} [11].

Definition 3.1. The mutual coherence of a matrix \mathbf{A} , $\mu(\mathbf{A})$, is the largest absolute correlation between any two columns of \mathbf{A} , i.e.,

$$\mu(\mathbf{A}) = \max_{i \neq j} \frac{|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2}, \quad (20)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

Intuitively, if two atoms in \mathbf{A} are highly correlated, then it will be difficult to distinguish their contributions to the measurements \mathbf{y} . In the extreme case when two atoms are completely coherent, it will be impossible to distinguish their contributions and thus impossible to recover the sparse signal \mathbf{x} . Therefore, to guarantee successful signal recovery, the mutual coherence $\mu(\mathbf{A})$ should be small. This is true, according to the following theorem.

Theorem 3.1 ([11]). Assume that $\|\mathbf{x}\|_0 \leq K$ for the true signal \mathbf{x} and $\mu < \frac{1}{2K-1}$. Then, \mathbf{x} is the unique solution of the ℓ_0 optimization and the BP problem.

Another theoretical guarantee is based on the restricted isometry property (RIP) that quantifies the correlations of the atoms in \mathbf{A} in a different manner and has been popular in the development of compressed sensing.

Definition 3.2 ([42]). The K -restricted isometry constant (RIC) of a matrix \mathbf{A} , $\delta_K(\mathbf{A})$, is the smallest number such that the inequality

$$(1 - \delta_K(\mathbf{A})) \|\mathbf{v}\|_2^2 \leq \|\mathbf{A}\mathbf{v}\|_2^2 \leq (1 + \delta_K(\mathbf{A})) \|\mathbf{v}\|_2^2$$

holds for all K -sparse vectors \mathbf{v} . \mathbf{A} is said to satisfy the K -RIP with constant $\delta_K(\mathbf{A})$ if $\delta_K(\mathbf{A}) < 1$.

The following theoretical guarantee is provided in [43].

Theorem 3.2 ([43]). Assume that $\|\mathbf{x}\|_0 \leq K$ for the true signal \mathbf{x} and $\delta_{2K} < \sqrt{2} - 1$. Then \mathbf{x} is the unique solution of the ℓ_0 optimization and the BP problem.

After the work [43], the RIP condition has been improved, e.g., to $\delta_{2K} < \frac{3}{4+\sqrt{6}}$ [44]. Other types of RIP conditions are also available, e.g., $\delta_K < 0.307$ in [45]. It is known that stronger results can be provided by using RIP as compared to the mutual coherence. But it is worth noting that, unlike the mutual coherence that can be easily computed given the matrix \mathbf{A} , the complexity of computing the RIC of \mathbf{A} may increase dramatically with the sparsity K .

In the presence of noise we can solve the following regularized optimization problem, usually known as the least absolute shrinkage and selection operator (LASSO) [46]:

$$\min_{\mathbf{x}} \lambda \|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2, \quad (21)$$

where $\lambda > 0$ is a regularization parameter, to be specified, or the basis pursuit denoising (BPDN) problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \text{ subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \eta, \quad (22)$$

where $\eta \geq \|\mathbf{e}\|_2$ is an upper bound on the noise energy. Note that (21) and (22) are equivalent for appropriate choices of λ and η , and that both degenerate to BP in the noiseless case by letting $\eta, \lambda \rightarrow 0$. Under RIP conditions similar to the above ones, it has been shown that the sparse signal \mathbf{x} can be stably reconstructed with the reconstruction error being proportional to the noise level [43].

Besides (21) and (22), another ℓ_1 optimization method for sparse recovery is the so-called square-root LASSO [47]:

$$\min_{\mathbf{x}} \tau \|\mathbf{x}\|_1 + \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2, \quad (23)$$

where $\tau > 0$ is a regularization parameter. Compared to the LASSO, for which the noise is usually assumed to be Gaussian and the regularization parameter λ is chosen proportional to the standard deviation of the noise, SR-LASSO requires a weaker assumption on the noise distribution and τ can be chosen as a constant that is independent of the noise level [47].

The ℓ_1 optimization problems in (19), (21), (22) and (23) are convex and are guaranteed to be solvable in a polynomial time; however, it is not easy to efficiently solve them in the case when the problem dimension is high since the ℓ_1 norm is not a smooth function. Significant progress has been made over the past decade to accelerate the computation. Examples include ℓ_1 -magic [48], interior-point method [49], conjugate gradient method [50], fixed-point continuation [51], Nesterov's smoothing technique with continuation (NESTA) [52, 53], ONE-L1 algorithms [54], alternating direction method of multipliers (ADMM) [55, 56] and so on.

3.1.3 ℓ_q Optimization

For a vector \mathbf{x} , the ℓ_q , $0 < q < 1$ (pseudo-)norm is defined as:

$$\|\mathbf{x}\|_q := \left(\sum_n |x_n|^q \right)^{\frac{1}{q}}, \quad (24)$$

which is a nonconvex relaxation of the ℓ_0 norm. Compared to (17) and (19), in the noiseless case, the ℓ_q optimization problem is given by:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_q^q, \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}, \quad (25)$$

where $\|\mathbf{x}\|_q^q$, instead of $\|\mathbf{x}\|_q$, is used for the convenience of algorithm development. Since the ℓ_q norm is a closer approximation to the ℓ_0 norm, compared to the ℓ_1 norm, it is expected that the ℓ_q optimization in (25) results in better performance than the BP. This is true, according to [29, 31]. Indeed, ℓ_q , $0 < q < 1$ optimization can exactly determine the true sparse signal under weaker RIP conditions than that for the BP. Note that the results above are applicable to the globally optimal solution to (25), whereas we can only guarantee convergence to a locally optimal solution in practice.

A well-known algorithm for ℓ_q optimization is the *focal underdetermined system solver* (FOCUSS) [26, 27]. FOCUSS is an iterative reweighted least squares method. In each iteration, FOCUSS solves the following weighted least squares problem:

$$\min_{\mathbf{x}} \sum_n w_n |x_n|^2, \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{y}, \quad (26)$$

where the weight coefficients $w_n := |x_n|^{q-2}$ are updated using the latest solution \mathbf{x} . Note that (26) can be solved in closed form and hence an iterative algorithm can be implemented with a proper initialization. This algorithm can be interpreted as a majorization-minimization (MM) algorithm that is guaranteed to converge to a local minimum.

In the presence of noise, the following regularized problem is considered in lieu of (21):

$$\min_{\mathbf{x}} \lambda \|\mathbf{x}\|_q^q + \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2, \quad (27)$$

where $\lambda > 0$ is a regularization parameter. A regularized FOCUSS algorithm for (27) was developed in [28] by using the same main idea as in FOCUSS. A difficult problem regarding (27) is the choice of the parameter λ . Although several heuristic methods for tuning this parameter were introduced in [28], to the best of our knowledge there have been no theoretical results on this aspect.

To bypass the parameter tuning problem, a maximum *a posterior* (MAP) estimation approach called SLIM (sparse learning via iterative minimization) was proposed in [32]. Assuming i.i.d. Gaussian noise with variance η and the following prior distribution for \mathbf{x} :

$$f(\mathbf{x}) \propto \prod_n e^{-\frac{2}{q}(|x_n|^q - 1)}, \quad (28)$$

SLIM computes the MAP estimate by solving the following ℓ_q optimization problem:

$$\min_{\mathbf{x}} M \log \eta + \eta^{-1} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \frac{2}{q} \|\mathbf{x}\|_q^q. \quad (29)$$

To locally solve (29), SLIM iteratively updates \mathbf{x} , as the regularized FOCUSS does. However, unlike FOCUSS, SLIM also iteratively updates the parameter η based on the latest solution \mathbf{x} . Once q is given, SLIM is hyper-parameter free. Note that (29) reduces to (27) for fixed η .

3.1.4 Maximum Likelihood Estimation (MLE)

MLE is another common approach to sparse estimation. In contrast to the convex relaxation and OMP, one advantage of MLE is that it does not require knowledge of the noise level or the sparsity level (the latter being often needed to choose λ in (21) properly). To derive it, assume that \mathbf{x} follows a multivariate Gaussian distribution with mean zero and covariance $\mathbf{P} = \text{diag}(\mathbf{p})$, where $p_n \geq 0$, $p = 1, \dots, \bar{N}$ (this can be viewed as a prior distribution that does not necessarily have to hold in practice). Also, assume i.i.d. Gaussian noise with variance σ . It follows from the data model in (16) that \mathbf{y} follows a Gaussian distribution with mean zero and covariance $\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma\mathbf{I}$. Consequently, the negative log-likelihood function associated with \mathbf{y} is given by

$$\mathcal{L}(\mathbf{p}, \sigma) = \log |\mathbf{R}| + \mathbf{y}^H \mathbf{R}^{-1} \mathbf{y}. \quad (30)$$

The parameters \mathbf{p} and σ can be estimated by minimizing \mathcal{L} :

$$\min_{\mathbf{p}, \sigma} \log |\mathbf{R}| + \mathbf{y}^H \mathbf{R}^{-1} \mathbf{y}. \quad (31)$$

Once \mathbf{p} and σ are solved for, the posterior distribution of the sparse signal \mathbf{x} can be obtained: it is a Gaussian distribution with mean and covariance given, respectively, by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} \mathbf{A}^H \mathbf{y}, \quad (32)$$

$$\boldsymbol{\Sigma} = (\mathbf{A}^H \mathbf{A} + \sigma \mathbf{P}^{-1})^{-1}. \quad (33)$$

The vector \mathbf{x} can be estimated as its posterior mean $\boldsymbol{\mu}$. In this process, the sparsity of \mathbf{x} is achieved by the fact that most of the entries of \mathbf{p} approach zero in practice. Theoretically, it can be shown that in the limiting noiseless case, the global optimizer to (31) coincides with that of ℓ_0 optimization [57].

The main difficulty of MLE is solving (31) in which the first term of the objective function, viz. $\log |\mathbf{R}|$, is a nonconvex (in fact, concave) function of (\mathbf{p}, σ) . Different approaches have been proposed, e.g., reweighted optimization [58] and sparse Bayesian learning (SBL) [57, 59, 60]. In [58], a majorization-minimization approach is adopted to linearize $\log |\mathbf{R}|$ at each iteration by its tangent plane $\text{Tr}(\mathbf{R}_j^{-1} \mathbf{R}) + \text{const}$ given the latest estimate \mathbf{R}_j . The resulting problem at each iteration is convex and solved using an algorithm called *sparse iterative covariance-based estimation* (SPICE) [58, 61–63] that will be introduced in Subsection 4.5.

The MLE has been interpreted from a different perspective within the framework of SBL or Bayesian compressed sensing. In particular, to achieve sparsity, a prior distribution is assumed for \mathbf{x} that promotes sparsity and is usually referred to as a sparse prior. In [59], for example, a Student's t -distribution is assumed

for \mathbf{x} that is constructed in a hierarchical manner: specifically, a Gaussian distribution as above at the first stage followed by a Gamma distribution for the inverse of the powers, p_n^{-1} , $n = 1, \dots, \bar{N}$ at the second stage. Interestingly, despite different approaches, the same objective function is obtained as in (31). To optimize (31), an expectation-maximization (EM) algorithm is adopted [64]. In the E-step, the posterior distribution of \mathbf{x} is computed, as mentioned previously, while in the M-step, \mathbf{p} and σ are updated as functions of the latest statistics of \mathbf{x} , viz. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The process is repeated and it guarantees a monotonic decrease of \mathcal{L} . Finally, we note that with other sparse priors for \mathbf{x} that may possess different sparsity promoting properties, the obtained objective function of SBL can be slightly different from that of the MLE in (31) (see, e.g., [65]).

3.2 Sparse Representation and DOA Estimation: the Link and the Gap

In this subsection we discuss the link and the gap between sparse representation and DOA estimation. By doing so, we can see the possibility and the main challenges of using the sparse representation techniques for DOA estimation. It has been mentioned that the underlying motivation of sparse representation is that the observed data \mathbf{y} can be well approximated in a lower-dimensional subspace. In fact, this is exactly the case in DOA estimation where the data snapshot $\mathbf{y}(t)$ is a linear combination of the steering vectors of the sources and the sparsity arises from the fact that there are less sources than sensors (note that for some special arrays and methods more sources than the sensors can be detected). By comparing the models in (1) and (16), it can be seen that the process of DOA estimation boils down to a sparse representation of the data snapshot with each DOA θ corresponding to one atom given by $\mathbf{a}(\theta)$. Therefore, it is possible to use sparse representation techniques in DOA estimation.

However, there exist major differences between the common sparse representation framework and DOA estimation. First, and most importantly, the dictionary in sparse representation usually contains a finite number of atoms while in the DOA estimation problem the parameters are continuously valued, which leads to infinitely many atoms. More concretely, the atoms in sparse representation are given by the columns of a matrix. But in DOA estimation each atom $\mathbf{a}(\theta)$ is parameterized by a continuous parameter θ .

Second, there are usually multiple snapshots in DOA estimation problems, in contrast to the single snapshot case in sparse representation. It is then crucial to exploit the temporal redundancy of the snapshots in DOA estimation since the number of antennas can be limited due to physical and other constraints. Typically, the number of antennas M is about $10 \sim 100$, while the number of snapshots L can be much larger.

Last but not least, the existing theoretical guarantees of the sparse representation techniques are usually derived using what is known as incoherence analysis, e.g., those based on the mutual coherence and RIP, in the sense that they are applicable only in the case of incoherent dictionaries. This means that such guarantees can hardly be applied to DOA estimation problems, in which the atoms are completely coherent. But this does not necessarily mean that satisfactory performance cannot be achieved in DOA estimation problems. Indeed, note that the success of sparse signal recovery is measured by the size of the reconstruction error of the sparse signal \mathbf{x} , and that a slight error in the support usually results in a large estimation error. But this is not true for DOA estimation where the estimation error is actually measured by the error of the support (and, therefore, a small estimation error of the support is acceptable).

The next three sections describe three different possibilities for dealing with the first gap—discrete versus continuous atoms—when applying sparse representation to DOA estimation. In each section, we will also discuss how the signal sparsity and the temporal redundancy of the multiple snapshots are exploited and what theoretical guarantees can be obtained.

4 On-Grid Sparse Methods

In this section we introduce the first class of sparse methods for DOA estimation, termed as on-grid sparse methods. These methods are developed by directly applying sparse representation and compressed sensing techniques to DOA estimation. To do so, the DOAs are assumed to lie on a prescribed grid so that the problem can be solved within the common framework of sparse representation. The main challenge then is how to exploit the temporal redundancy of multiple snapshots.

In the following we first introduce the data model that we will use throughout this section. Then, we present several formulations and algorithms for DOA estimation within the on-grid framework, including $\ell_{2,q}$ optimization methods with $0 \leq q \leq 1$, SBL and SPICE. Guidelines for grid selection will also be provided.

4.1 Data Model

To fill the gap between continuous DOA estimation and discrete sparse representation, it is simply assumed by the on-grid sparse methods that the continuous DOA domain \mathcal{D}_θ can be replaced by a *given* set of grid points

$$\bar{\boldsymbol{\theta}} := \{\bar{\theta}_1, \dots, \bar{\theta}_{\bar{N}}\}, \quad (34)$$

where $\bar{N} \gg M$ is the grid size. This means that the candidate DOAs can only take values in $\bar{\boldsymbol{\theta}}$, which results in the following $M \times \bar{N}$ dictionary matrix

$$\mathbf{A} := \mathbf{A}(\bar{\boldsymbol{\theta}}) = [\mathbf{a}(\bar{\theta}_1), \dots, \mathbf{a}(\bar{\theta}_{\bar{N}})]. \quad (35)$$

It follows that the data model in (2) for DOA estimation can be equivalently written as:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}, \quad (36)$$

where $\mathbf{X} = [\mathbf{x}(1), \dots, \mathbf{x}(L)]$ is an $\bar{N} \times L$ matrix in which each column $\mathbf{x}(t)$ is an augmented version of the source signal $\mathbf{s}(t)$ and is defined by:

$$x_n(t) = \begin{cases} s_k(t), & \text{if } \bar{\theta}_n = \theta_k; \\ 0, & \text{otherwise,} \end{cases} \quad n = 1, \dots, \bar{N}, t = 1, \dots, L. \quad (37)$$

It can be seen that for each t , $\mathbf{x}(t)$ contains only K nonzero entries, whose locations correspond to the K DOAs, and therefore it is a sparse vector as $K \ll \bar{N}$. Moreover, $\mathbf{x}(t)$, $t = 1, \dots, L$ are jointly sparse in the sense that they share the same support. Alternatively, we can say that \mathbf{X} is row-sparse in the sense that it contains only a few nonzero rows.

By means of the data model in (36), the DOA estimation problem is transformed into a sparse signal recovery problem. The DOAs are encoded in the support of the sparse vectors $\mathbf{x}(t)$, $t = 1, \dots, L$ and therefore, we only need to recover this support from which the estimated DOAs can be retrieved.

The key and only difference between (36) and (16) is that the former contains multiple data snapshots that are also referred to as multiple measurement vectors (MMVs). In the case of a single snapshot with $L = 1$ (i.e., single measurement vector (SMV)), the sparse representation techniques can be readily applied to DOA estimation. In the case of multiple snapshots, the main difficulty consists in exploiting the temporal redundancy of the snapshots—the joint sparsity of the columns of \mathbf{X} —for possibly improved performance. Since the MMV data model in (36) is quite general, extensive studies have been performed for the joint sparse signal recovery problem (see, e.g., [18, 19, 62, 66–78]). We only discuss some of them in the ensuing subsections.

Before proceeding to the on-grid sparse methods, we make some comments on the data model in (36). Note that the set of grid points $\bar{\boldsymbol{\theta}}$ needs to be fixed *a priori* so that the dictionary \mathbf{A} is known, which is

required in the sparse signal recovery process. Consequently, there is no guarantee that the true DOAs lie on the grid $\bar{\theta}$; in fact, this fails with probability one, resulting in the grid mismatch problem [79, 80]. To ensure at least that the true DOAs are *close* to the grid points, in practice the grid needs to be dense enough (with $\bar{N} \gg M$). Therefore, (36) can be viewed as a zeroth-order approximation of the true data model in (2) and the noise term \mathbf{E} in (36) may also comprise the approximation error besides the true noise in (2).

4.2 $\ell_{2,0}$ Optimization

We first discuss how the joint sparsity can be exploited for sparse recovery. We start with the definition of sparsity for the row-sparse matrix \mathbf{X} . Since each row of \mathbf{X} corresponds to one potential source, it is natural to define the sparsity as the number of nonzero rows of \mathbf{X} , which is usually expressed as the following $\ell_{2,0}$ norm (see, e.g., [18, 75]):

$$\|\mathbf{X}\|_{2,0} := \#\{n : \|\mathbf{X}_n\|_2 > 0\} = \#\{n : \mathbf{X}_n \neq \mathbf{0}\}, \quad (38)$$

where \mathbf{X}_n denotes the n th row of \mathbf{X} . Note that in (38) the ℓ_2 norm can in fact be replaced by any other norm. Following from the ℓ_0 optimization in the single snapshot case, the following $\ell_{2,0}$ optimization can be proposed in the absence of noise:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{2,0}, \text{ subject to } \mathbf{Y} = \mathbf{A}\mathbf{X}. \quad (39)$$

Suppose the optimal solution, denoted by $\widehat{\mathbf{X}}$, can be obtained. Then, the DOAs can be retrieved from the row-support of $\widehat{\mathbf{X}}$.

To realize the potential advantage of using the joint sparsity of the snapshots, consider the following result.

Theorem 4.1 ([18]). *The true matrix \mathbf{X} is the unique solution to (39) if*

$$\|\mathbf{X}\|_{2,0} < \frac{\text{spark}(\mathbf{A}) - 1 + \text{rank}(\mathbf{Y})}{2}. \quad (40)$$

Note that the condition in (40) is very similar to that in (10) required to guarantee parameter identifiability for DOA estimation. By Theorem 4.1, the number of recoverable DOAs can be increased in general by collecting more snapshots since then $\text{rank}(\mathbf{Y})$ increases. The only exception happens in the case when the data snapshots $\mathbf{y}(t)$, $t = 1, \dots, L$ are identical up to scaling factors. Unfortunately, similar to the single snapshot case, the above $\ell_{2,0}$ optimization problem is NP-hard to solve.

4.3 Convex Relaxation

4.3.1 $\ell_{2,1}$ Optimization

The tightest convex relaxation of the $\ell_{2,0}$ norm is given by the $\ell_{2,1}$ norm that is defined as:

$$\|\mathbf{X}\|_{2,1} := \sum_n \|\mathbf{X}_n\|_2. \quad (41)$$

Though in the definition in (38) the ℓ_2 norm in the $\ell_{2,0}$ norm can be replaced by other norms, its use is important in the $\ell_{2,1}$ norm. Based on (39), the following $\ell_{2,1}$ optimization problem is proposed in the absence of noise [66, 67]:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{2,1}, \text{ subject to } \mathbf{Y} = \mathbf{A}\mathbf{X}. \quad (42)$$

As reported in the literature, the performance of $\ell_{2,1}$ optimization approach can be generally improved by increasing the number of measurement vectors. Theoretically, this can be shown to be true under the assumption that the jointly sparse signals are randomly drawn such that the rows of the source signals \mathbf{S}_k , $k = 1, \dots, K$ are at general positions [74]. It is worth noting that the theoretical guarantee cannot be improved without assumptions on the source signals. To see this, consider the case when the columns of \mathbf{S} are identical up to scaling factors. Then, acquiring more snapshots does not provide useful information for DOA estimation. In this respect, the result of [74] can be referred to as *average case* analysis while those accounting for the aforementioned extreme case can be called *worst case* analysis.

In parallel to (21) and (22), in the presence of noise we can solve the LASSO problem:

$$\min_{\mathbf{X}} \lambda \|\mathbf{X}\|_{2,1} + \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_{\text{F}}^2 \quad (43)$$

where $\lambda > 0$ is a regularization parameter (to be specified), or the BPDN problem:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{2,1}, \text{ subject to } \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_{\text{F}} \leq \eta, \quad (44)$$

where $\eta \geq \|\mathbf{E}\|_2$ is an upper bound on the noise energy. Note that it is generally difficult to choose λ in (43). Given the noise variance, results on choosing λ have recently been provided in [81–83] in the case of ULA and SLA. Readers are referred to Section 6 for details.

Finally, we note that most, if not all, of the computational approaches to ℓ_1 optimization, e.g., those mentioned in Subsection 3.1.2, can be easily extended to deal with $\ell_{2,1}$ optimization in the case of multiple snapshots. Once \mathbf{X} is solved for, we can form a power spectrum by computing the power of each row of \mathbf{X} from which the estimated DOAs can be obtained.

4.3.2 Dimensionality Reduction via $\ell_{2,1}$ -SVD

In DOA estimation applications the number of snapshots L can be large, which can significantly increase the computational workload of $\ell_{2,1}$ optimization. In the case of $L > K$ a dimensionality reduction technique was proposed in [67] inspired by the conventional subspace methods, e.g., MUSIC. In particular, suppose there is no noise; then the data snapshots \mathbf{Y} lie in a K -dimensional subspace. In the presence of noise, therefore, we can decompose \mathbf{Y} into the signal and noise subspaces, keep the signal subspace and use it in (42)-(44) in lieu of \mathbf{Y} . Mathematically, we compute the singular value decomposition (SVD)

$$\mathbf{Y} = \mathbf{U}\mathbf{L}\mathbf{V}^H. \quad (45)$$

Define a reduced $M \times K$ dimensional data matrix

$$\mathbf{Y}_{\text{SV}} = \mathbf{U}\mathbf{L}\mathbf{D}_K^T = \mathbf{Y}\mathbf{V}\mathbf{D}_K^T \quad (46)$$

that contains most of the signal power, where $\mathbf{D}_K := [\mathbf{I}_K, \mathbf{0}]$ with \mathbf{I}_K being an identity matrix of order K . Also let $\mathbf{X}_{\text{SV}} = \mathbf{X}\mathbf{V}\mathbf{D}_K^T$ and $\mathbf{E}_{\text{SV}} = \mathbf{E}\mathbf{V}\mathbf{D}_K^T$. Using these notations we can write a new data model:

$$\mathbf{Y}_{\text{SV}} = \mathbf{A}\mathbf{X}_{\text{SV}} + \mathbf{E}_{\text{SV}}. \quad (47)$$

Note that (47) is in exactly the same form as (36) but with reduced dimensionality. So similar $\ell_{2,1}$ optimization problems can be formulated as (43) and (44), which are referred to as $\ell_{2,1}$ -SVD.

The following comments on $\ell_{2,1}$ -SVD are in order. Note that the true source number K has been known to obtain \mathbf{Y}_{SV} . However, $\ell_{2,1}$ -SVD is not very sensitive to this choice and therefore an appropriate estimate of K is sufficient in practice [67]. Nevertheless, parameter tuning remains a difficult problem for $\ell_{2,1}$ -SVD (λ and η in (43) and (44)). Though some solutions have been proposed for the standard $\ell_{2,1}$ optimization

methods in (43) and (44), given the noise level, they cannot be applied to $\ell_{2,1}$ -SVD due to the change in the data structure. Regarding this aspect, it is somehow hard to compare the DOA estimation performances of the standard $\ell_{2,1}$ optimization and $\ell_{2,1}$ -SVD; though it is argued in [67] that, as compared to the standard form, $\ell_{2,1}$ -SVD can improve the robustness to noise by keeping only the signal subspace.

4.3.3 Another Dimensionality Reduction Technique

We present here another dimensionality reduction technique that reduces the number of snapshots from L to M and has the same performance as the original $\ell_{2,1}$ optimization. The technique was proposed in [84], inspired by a similar technique used for the gridless sparse methods (see Section 6). For convenience, we introduce it following the idea of $\ell_{2,1}$ -SVD. In $\ell_{2,1}$ -SVD the number of snapshots is reduced from L to K by keeping only the K -dimensional signal subspace; in contrast the technique in [84] suggests keeping both the signal and noise subspaces. To be specific, suppose that $L > M$ and \mathbf{Y} has rank $r \leq M$ (note that typically $r = M$ in the presence of noise). Then, given the SVD in (45), we retain a reduced $M \times r$ dimensional data matrix

$$\mathbf{Y}_{\text{DR}} = \mathbf{U} \mathbf{L} \mathbf{D}_r^T = \mathbf{Y} \mathbf{V} \mathbf{D}_r^T \quad (48)$$

that preserves all of the data power since \mathbf{Y} has only r nonzero singular values, where \mathbf{D}_r is defined similarly to \mathbf{D}_K . We similarly define $\mathbf{X}_{\text{DR}} = \mathbf{X} \mathbf{V} \mathbf{D}_r^T$ and $\mathbf{E}_{\text{DR}} = \mathbf{E} \mathbf{V} \mathbf{D}_r^T$ to obtain the data model

$$\mathbf{Y}_{\text{DR}} = \mathbf{A} \mathbf{X}_{\text{DR}} + \mathbf{E}_{\text{DR}}. \quad (49)$$

For the LASSO problem, as an example, it can be shown that *equivalent* solutions can be obtained before and after the dimensionality reduction. To be specific, if $\widehat{\mathbf{X}}_{\text{DR}}$ is the solution to the following LASSO problem:

$$\min_{\mathbf{X}_{\text{DR}}} \lambda \|\mathbf{X}_{\text{DR}}\|_{2,1} + \frac{1}{2} \|\mathbf{A} \mathbf{X}_{\text{DR}} - \mathbf{Y}_{\text{DR}}\|_{\text{F}}^2, \quad (50)$$

then $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}_{\text{DR}} \mathbf{D}_r \mathbf{V}^H$ is the solution to (43). $\widehat{\mathbf{X}}_{\text{DR}}$ and $\widehat{\mathbf{X}}$ are equivalent in the sense that the corresponding rows of $\widehat{\mathbf{X}}_{\text{DR}}$ and $\widehat{\mathbf{X}}$ have the same power, resulting in identical power spectra (to see this, note that $\widehat{\mathbf{X}} \widehat{\mathbf{X}}^H = \widehat{\mathbf{X}}_{\text{DR}} \mathbf{D}_r \mathbf{V}^H \mathbf{V} \mathbf{D}_r^T \widehat{\mathbf{X}}_{\text{DR}}^H = \widehat{\mathbf{X}}_{\text{DR}} \widehat{\mathbf{X}}_{\text{DR}}^H$, whose diagonal contains the powers of the rows).

We next prove the above result. To do so, for any \mathbf{X} , split $\mathbf{X} \mathbf{V}$ into two parts: $\mathbf{X} \mathbf{V} = [\mathbf{X}_{\text{DR}}, \mathbf{X}_2]$. Note that $\mathbf{Y} \mathbf{V} = [\mathbf{Y}_{\text{DR}}, \mathbf{0}]$. By the fact that \mathbf{V} is a unitary matrix, it can be easily shown that

$$\|\mathbf{X}\|_{2,1} = \|\mathbf{X} \mathbf{V}\|_{2,1} = \|[\mathbf{X}_{\text{DR}}, \mathbf{X}_2]\|_{2,1}, \quad (51)$$

$$\|\mathbf{A} \mathbf{X} - \mathbf{Y}\|_{\text{F}}^2 = \|\mathbf{A} \mathbf{X} \mathbf{V} - \mathbf{Y} \mathbf{V}\|_{\text{F}}^2 = \|\mathbf{A} \mathbf{X}_{\text{DR}} - \mathbf{Y}_{\text{DR}}\|_{\text{F}}^2 + \|\mathbf{A} \mathbf{X}_2\|_{\text{F}}^2. \quad (52)$$

It immediately follows that

$$\lambda \|\mathbf{X}\|_{2,1} + \frac{1}{2} \|\mathbf{A} \mathbf{X} - \mathbf{Y}\|_{\text{F}}^2 \geq \lambda \|\mathbf{X}_{\text{DR}}\|_{2,1} + \frac{1}{2} \|\mathbf{A} \mathbf{X}_{\text{DR}} - \mathbf{Y}_{\text{DR}}\|_{\text{F}}^2 \quad (53)$$

and the equality holds if and only if $\mathbf{X}_2 = \mathbf{0}$, or equivalently, $\mathbf{X} = \mathbf{X}_{\text{DR}} \mathbf{D}_r \mathbf{V}^H$. We can obtain the stated result by minimizing both sides of (53) with respect to \mathbf{X} .

Note that the above result also holds if \mathbf{Y}_{DR} is replaced by any full-column-rank matrix $\widetilde{\mathbf{Y}}$ satisfying $\widetilde{\mathbf{Y}} \widetilde{\mathbf{Y}}^H = \mathbf{Y} \mathbf{Y}^H$, since there always exists a unitary matrix \mathbf{V} such that $\widetilde{\mathbf{Y}} = \mathbf{Y} \mathbf{V} \mathbf{D}_r^T$ as for \mathbf{Y}_{DR} . Therefore, the SVD of the $M \times L$ dimensional data matrix \mathbf{Y} , which can be computationally expensive in the case of $L \gg M$, can be replaced by the Cholesky decomposition or the eigenvalue decomposition of the $M \times M$ matrix $\mathbf{Y} \mathbf{Y}^H$ (which is the sample data covariance matrix up to a scaling factor). Another fact that makes this dimensional reduction technique superior to $\ell_{2,1}$ -SVD is that the parameter λ or η can be tuned as in the original $\ell_{2,1}$ optimization, for which solutions are available if the noise level is given.

4.4 $\ell_{2,q}$ Optimization

Corresponding to the ℓ_q , $0 < q < 1$ norm considered in Subsection 3.1.3, we can define the $\ell_{2,q}$ norm to exploit the joint sparsity in \mathbf{X} as:

$$\|\mathbf{X}\|_{2,q} = \left(\sum_n \|\mathbf{X}_n\|_2^q \right)^{\frac{1}{q}}, \quad (54)$$

which is a nonconvex relaxation of the $\ell_{2,0}$ norm. In lieu of (42) and (43), in the noiseless case, we can solve the following equality constrained problem:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{2,q}^q, \text{ subject to } \mathbf{A}\mathbf{X} = \mathbf{Y}, \quad (55)$$

or the following regularized form in the noisy case:

$$\min_{\mathbf{X}} \lambda \|\mathbf{X}\|_{2,q}^q + \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_F^2. \quad (56)$$

To locally solve (55) and (56), the FOCUSS algorithm was extended in [66] to this multiple snapshot case to obtain M-FOCUSS. For (55), as in the single snapshot case, M-FOCUSS solves the following weighted least squares problem in each iteration:

$$\min_{\mathbf{X}} \sum_n w_n \|\mathbf{X}\|_2^2, \text{ subject to } \mathbf{A}\mathbf{X} = \mathbf{Y}, \quad (57)$$

where the weight coefficients $w_n := \|\mathbf{X}\|_2^{q-2}$ are updated based on the latest solution \mathbf{X} . Since (57) can be solved in closed form, an iterative algorithm can be implemented. Note that (56) can be similarly solved as (55).

To circumvent the need for tuning the regularization parameter λ in (56), SLIM [32] can be extended to this multiple snapshot case. Under the assumption of i.i.d. Gaussian noise with variance η and assuming that \mathbf{X} follows a prior distribution with the pdf given by

$$f(\mathbf{X}) \propto \prod_n e^{-\frac{2}{q}(\|\mathbf{X}_n\|_2^q - 1)}, \quad (58)$$

SLIM computes the MAP estimator of \mathbf{X} by solving the following $\ell_{2,q}$ optimization problem:

$$\min_{\mathbf{x}} ML \log \eta + \eta^{-1} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_F^2 + \frac{2}{q} \|\mathbf{X}\|_{2,q}^q. \quad (59)$$

Using a reweighting technique similar to that in M-FOCUSS, we can iteratively update \mathbf{X} and η in closed form and obtain the multiple snapshot version of SLIM. Finally, note that the dimensionality reduction technique presented in Subsection 4.3.3 can also be applied to the $\ell_{2,q}$ optimization problems in (55), (56) and (59) for algorithm acceleration [84].

4.5 Sparse Iterative Covariance-based Estimation (SPICE)

4.5.1 Generalized Least Squares

To introduce SPICE, we first present the so-called generalized least squares method. To derive it, we need some statistical assumptions on the sources \mathbf{X} and the noise \mathbf{E} . We assume that $\{\mathbf{x}(1), \dots, \mathbf{x}(L), \mathbf{e}(1), \dots, \mathbf{e}(L)\}$ are uncorrelated with one another and

$$\mathbb{E} \mathbf{e}(t) \mathbf{e}^H(t) = \sigma \mathbf{I}, \quad (60)$$

$$\mathbb{E} \mathbf{x}(t) \mathbf{x}^H(t) = \mathbf{P} := \text{diag}(\mathbf{p}), \quad t = 1, \dots, L, \quad (61)$$

where $\sigma \geq 0$ and $p_n \geq 0$, $n = 1, \dots, \bar{N}$ are the parameters of interest (note that the following derivations also apply to the case of heteroscedastic noise with $\mathbb{E}e(t)e^H(t) = \text{diag}(\sigma_1, \dots, \sigma_M)$ with no or minor modifications). It follows that the snapshots $\{\mathbf{y}(1), \dots, \mathbf{y}(L)\}$ are uncorrelated with one another and have the following covariance matrix:

$$\mathbf{R}(\mathbf{p}, \sigma) = \mathbb{E}\mathbf{y}(t)\mathbf{y}^H(t) = \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma\mathbf{I} := \mathbf{A}'\mathbf{P}'\mathbf{A}'^H, \quad (62)$$

where $\mathbf{A}' := [\mathbf{A}, \mathbf{I}]$ and $\mathbf{P}' := \text{diag}(\mathbf{P}, \sigma\mathbf{I})$. Note that \mathbf{R} is linear in (\mathbf{p}, σ) . Let $\tilde{\mathbf{R}} = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H$ be the sample covariance matrix. Given $\tilde{\mathbf{R}}$, to estimate \mathbf{R} (in fact, the parameters \mathbf{p} and σ therein), we consider the generalized least squares method. First, we vectorize $\tilde{\mathbf{R}}$ and let $\tilde{\mathbf{r}} = \text{vec}(\tilde{\mathbf{R}})$ and $\mathbf{r} = \text{vec}(\mathbf{R})$. Since $\tilde{\mathbf{R}}$ is an unbiased estimate of the data covariance matrix, it holds that

$$\mathbb{E}\tilde{\mathbf{r}} = \mathbf{r}. \quad (63)$$

Moreover, we can calculate the covariance matrix of $\tilde{\mathbf{r}}$, which is given by (see, e.g., [85])

$$\text{Cov}(\tilde{\mathbf{r}}) = \frac{1}{L}\mathbf{R}^T \otimes \mathbf{R}, \quad (64)$$

where \otimes denotes the Kronecker product. In the generalized least squares method we minimize the following criterion [85, 86]:

$$\begin{aligned} & \frac{1}{L} (\tilde{\mathbf{r}} - \mathbb{E}\tilde{\mathbf{r}})^H \text{Cov}^{-1}(\tilde{\mathbf{r}}) (\tilde{\mathbf{r}} - \mathbb{E}\tilde{\mathbf{r}}) \\ &= (\tilde{\mathbf{r}} - \mathbf{r})^H [\mathbf{R}^{-T} \otimes \mathbf{R}^{-1}] (\tilde{\mathbf{r}} - \mathbf{r}) \\ &= \text{vec}^H(\tilde{\mathbf{R}} - \mathbf{R}) [\mathbf{R}^{-T} \otimes \mathbf{R}^{-1}] \text{vec}(\tilde{\mathbf{R}} - \mathbf{R}) \\ &= \text{vec}^H(\tilde{\mathbf{R}} - \mathbf{R}) \text{vec}\left\{\mathbf{R}^{-1}(\tilde{\mathbf{R}} - \mathbf{R})\mathbf{R}^{-1}\right\} \\ &= \text{Tr}\left\{(\tilde{\mathbf{R}} - \mathbf{R})\mathbf{R}^{-1}(\tilde{\mathbf{R}} - \mathbf{R})\mathbf{R}^{-1}\right\} \\ &= \left\|\mathbf{R}^{-\frac{1}{2}}(\tilde{\mathbf{R}} - \mathbf{R})\mathbf{R}^{-\frac{1}{2}}\right\|_F^2. \end{aligned} \quad (65)$$

The criterion in (65) has good statistical properties; for example, under certain conditions it provides a large-snapshot maximum likelihood (ML) estimator of the parameters (\mathbf{p}, σ) of interest. Unfortunately, (65) is nonconvex in \mathbf{R} and hence nonconvex in (\mathbf{p}, σ) . Therefore, there is no guarantee that it can be globally minimized.

Inspired by (65), the following convex criterion was proposed in [87]:

$$\left\|\tilde{\mathbf{R}}^{-\frac{1}{2}}(\tilde{\mathbf{R}} - \mathbf{R})\tilde{\mathbf{R}}^{-\frac{1}{2}}\right\|_F^2, \quad (66)$$

in which $\text{Cov}(\tilde{\mathbf{r}})$ in (64) is replaced by its consistent estimate, viz. $\frac{1}{L}\tilde{\mathbf{R}}^T \otimes \tilde{\mathbf{R}}$. The resulting estimator remains a large-snapshot ML estimator. But it is only usable in the case of $L \geq M$ when $\tilde{\mathbf{R}}$ is nonsingular. The SPICE approach, which is discussed next, relies on (65) or (66) (see below for details).

4.5.2 SPICE

The SPICE algorithm has been proposed and studied in [58, 61–63]. In SPICE, the following covariance fitting criterion is adopted in the case of $L \geq M$ whenever $\tilde{\mathbf{R}}$ is nonsingular:

$$h_1 = \left\|\mathbf{R}^{-\frac{1}{2}}(\tilde{\mathbf{R}} - \mathbf{R})\tilde{\mathbf{R}}^{-\frac{1}{2}}\right\|_F^2. \quad (67)$$

In the case of $L < M$, in which $\tilde{\mathbf{R}}$ is singular, the following criterion is used instead:

$$h_2 = \left\| \mathbf{R}^{-\frac{1}{2}} \left(\tilde{\mathbf{R}} - \mathbf{R} \right) \right\|_{\text{F}}^2. \quad (68)$$

A simple calculation shows that

$$\begin{aligned} h_1 &= \text{Tr} \left(\mathbf{R}^{-1} \tilde{\mathbf{R}} \right) + \text{Tr} \left(\tilde{\mathbf{R}}^{-1} \mathbf{R} \right) - 2M \\ &= \text{Tr} \left(\tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}} \right) + \sum_{n=1}^{\bar{N}} \left(\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n \right) p_n + \text{Tr} \left(\tilde{\mathbf{R}}^{-1} \right) \sigma - 2M. \end{aligned} \quad (69)$$

It follows that the optimization problem of SPICE based on h_1 can be equivalently formulated as:

$$\min_{\mathbf{p} \succeq \mathbf{0}, \sigma > 0} \text{Tr} \left(\tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}} \right) + \sum_{n=1}^{\bar{N}} \left(\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n \right) p_n + \text{Tr} \left(\tilde{\mathbf{R}}^{-1} \right) \sigma. \quad (70)$$

Note that the first term of the above objective function can be written as:

$$\text{Tr} \left(\tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}} \right) = \min \text{Tr}(\mathbf{X}), \text{ subject to } \begin{bmatrix} \mathbf{X} & \tilde{\mathbf{R}}^{\frac{1}{2}} \\ \tilde{\mathbf{R}}^{\frac{1}{2}} & \mathbf{R} \end{bmatrix} \succeq \mathbf{0} \quad (71)$$

and hence it is convex in \mathbf{R} as well as in (\mathbf{p}, σ) . It follows that h_1 is convex in (\mathbf{p}, σ) . Similarly, it holds for h_2 that

$$\begin{aligned} h_2 &= \text{Tr} \left(\mathbf{R}^{-1} \tilde{\mathbf{R}}^2 \right) + \text{Tr}(\mathbf{R}) - 2\text{Tr}(\tilde{\mathbf{R}}) \\ &= \text{Tr} \left(\tilde{\mathbf{R}} \mathbf{R}^{-1} \tilde{\mathbf{R}} \right) + \sum_{n=1}^{\bar{N}} \|\mathbf{a}_n\|_2^2 p_n + M\sigma - 2\text{Tr}(\tilde{\mathbf{R}}). \end{aligned} \quad (72)$$

The resulting optimization problem is given by:

$$\min_{\mathbf{p} \succeq \mathbf{0}, \sigma > 0} \text{Tr} \left(\tilde{\mathbf{R}} \mathbf{R}^{-1} \tilde{\mathbf{R}} \right) + \sum_{n=1}^{\bar{N}} \|\mathbf{a}_n\|_2^2 p_n + M\sigma, \quad (73)$$

which is in a form similar to (70) and therefore is convex as well. Although both (70) and (73) can be cast as second order cone programs (SOCP) or semidefinite programs (SDP) (as shown above), for which standard solvers are available, they cannot be practically solved based on these formulations due to the curse of dimensionality (note that \bar{N} can be very large).

We now introduce the SPICE algorithm to cope with the aforementioned computational problems. We focus on the case of $L \geq M$ but similar results also hold in the case of $L < M$. The main result that underlies SPICE is the following reformulation (see, e.g., [62]):

$$\text{Tr} \left(\tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}} \right) = \min_{\mathbf{C}} \text{Tr}(\mathbf{C}^H \mathbf{P}'^{-1} \mathbf{C}), \text{ subject to } \mathbf{A}' \mathbf{C} = \tilde{\mathbf{R}}^{\frac{1}{2}} \quad (74)$$

and showing that the solution of \mathbf{C} is given by

$$\mathbf{C} = \mathbf{P}' \mathbf{A}'^H \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}}. \quad (75)$$

Inserting (74) into (70), we see that the minimization of h_1 can be equivalently written as:

$$\min_{\mathbf{C}, \mathbf{p} \succeq \mathbf{0}, \sigma > 0} \text{Tr}(\mathbf{C}^H \mathbf{P}'^{-1} \mathbf{C}) + \sum_{n=1}^{\bar{N}} \left(\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n \right) p_n + \text{Tr}(\tilde{\mathbf{R}}^{-1}) \sigma, \quad (76)$$

subject to $\mathbf{A}' \mathbf{C} = \tilde{\mathbf{R}}^{\frac{1}{2}}$.

Based on (76), the SPICE algorithm is derived by iteratively solving for \mathbf{C} and for (\mathbf{p}, σ) . First, \mathbf{p} and σ are initialized using, e.g., the conventional beamformer. Then, \mathbf{C} is updated using (75) with the latest estimates of \mathbf{p} and σ . After that, we update \mathbf{p} and σ by fixing \mathbf{C} and repeat the process until convergence. Note that (\mathbf{p}, σ) can also be determined in closed form, for fixed \mathbf{C} . To see this, observe that

$$\text{Tr}(\mathbf{C}^H \mathbf{P}'^{-1} \mathbf{C}) = \sum_{n=1}^{\bar{N}} \frac{\|\mathbf{C}_n\|_2^2}{p_n} + \frac{\sum_{n=\bar{N}+1}^{\bar{N}+M} \|\mathbf{C}_n\|_2^2}{\sigma}, \quad (77)$$

where \mathbf{C}_n denotes the n th row of \mathbf{C} . Inserting (77) in (76), the solutions p_n , $n = 1, \dots, \bar{N}$ and σ can be obtained as:

$$p_n = \frac{\|\mathbf{C}_n\|_2}{\sqrt{\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n}}, \quad n = 1, \dots, \bar{N}, \quad (78)$$

$$\sigma = \sqrt{\frac{\sum_{n=\bar{N}+1}^{\bar{N}+M} \|\mathbf{C}_n\|_2^2}{\text{Tr}(\tilde{\mathbf{R}}^{-1})}}. \quad (79)$$

Since the problem is convex and the objective function is monotonically decreasing in the iterative process, the SPICE algorithm is expected to converge to the global minimum. SPICE has a per-iteration computational complexity of $O(\bar{N}M^2)$ given that $\bar{N} > M$. Note also that, as compared to the original SPICE algorithm in [61, 62], a certain normalization step of the power estimates is removed here to avoid a global scaling ambiguity of the final power estimates (see also [58]).

We next discuss how the signal sparsity and joint sparsity are exploited in SPICE. Inserting (78) and (79) into (76), we see that the SPICE problem is equivalent to:

$$\min_{\mathbf{C}} \sum_{n=1}^{\bar{N}} \sqrt{\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n} \|\mathbf{C}_n\|_2 + \sqrt{\text{Tr}(\tilde{\mathbf{R}}^{-1}) \sum_{n=\bar{N}+1}^{\bar{N}+M} \|\mathbf{C}_n\|_2^2}, \quad (80)$$

subject to $\mathbf{A}' \mathbf{C} = \tilde{\mathbf{R}}^{\frac{1}{2}}$.

Note that the first term of the objective function in (80) is nothing but a weighted sum of the ℓ_2 norm of the first \bar{N} rows of \mathbf{C} (a.k.a. a weighted $\ell_{2,1}$ norm) and thus promotes the row-sparsity of \mathbf{C} . Therefore, it is expected that most of $\|\mathbf{C}_n\|_2$, $n = 1, \dots, \bar{N}$ will be equal to zero. This together with (78) implies that most of p_n , $n = 1, \dots, \bar{N}$ will be zero and so sparsity is achieved. The joint sparsity is achieved by the assumption that the entries in each row of \mathbf{X} have identical variance p_n .

Finally, note that SPICE is related to the square-root LASSO in the single snapshot case. In particular, it was shown in [88, 89] that the SPICE problem in (73) is equivalent to

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2, \quad (81)$$

which is nothing but the square-root LASSO in (23) with $\tau = 1$.

4.6 Maximum Likelihood Estimation

The joint sparsity can also be exploited in the MLE, in a similar way as in SPICE. Assume that $\mathbf{x}(t)$, $t = 1, \dots, L$ are i.i.d. multivariate Gaussian distributed with mean zero and covariance $\mathbf{P} = \text{diag}(\mathbf{p})$. Also assume i.i.d. Gaussian noise with variance σ and that \mathbf{X} and \mathbf{E} are independent. Then, we have that the data snapshots $\mathbf{y}(t)$, $t = 1, \dots, L$ are i.i.d. Gaussian distributed with mean zero and covariance $\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma\mathbf{I}$. The negative log-likelihood function associated with \mathbf{Y} is therefore given by

$$\mathcal{L}(\mathbf{p}, \sigma) = \log |\mathbf{R}| + \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}) \quad (82)$$

where $\tilde{\mathbf{R}}$ is the sample covariance matrix as defined in the preceding subsection. It follows that the parameters \mathbf{p} and σ can be estimated by solving the problem:

$$\min_{\mathbf{p}, \sigma} \log |\mathbf{R}| + \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}). \quad (83)$$

Owing to the analogy between (83) and its single snapshot counterpart (see (31)), it should come as no surprise that the algorithms developed for the single snapshot case can also be applied to the multiple snapshot case with minor modifications. As an example, using a similar MM procedure, LIKES [58] can be extended to this multiple snapshot case.

The multiple snapshot MLE has been studied within the framework of SBL or Bayesian compressed sensing (see, e.g., [72, 90–92]). To exploit the joint sparsity of $\mathbf{x}(t)$, $t = 1, \dots, L$, an identical sparse prior was assumed for all of them. The EM algorithm can also be used to perform parameter estimation via minimizing the objective in (83).

4.7 Remarks on Grid Selection

Based on the on-grid data model in (36), we have introduced several sparse optimization methods for DOA estimation in the preceding subsections. While we have focused on how the temporal redundancy or joint sparsity of the snapshots can be exploited, a major problem that remains unresolved is grid selection, i.e., the selection of the grid points $\bar{\boldsymbol{\theta}}$ in the data model (36). Since discrete grid points are used to approximate the continuous DOA domain, intuitively, the grid should be chosen as fine as possible to improve the approximation accuracy. However, this can be problematic in two respects. Theoretically, a dense grid results in highly coherent atoms and hence few DOAs can be estimated according to the analysis based on the mutual coherence and RIP. Moreover, a too dense grid is not acceptable from an algorithmic viewpoint, since it will dramatically increase the computational complexity of an algorithm and also might cause slow convergence and numerical instability due to nearly identical adjacent atoms [93, 94].

To overcome the theoretical bottleneck mentioned above, the so-called coherence-inhibiting techniques have been proposed and incorporated in the existing sparse optimization methods to avoid solutions with closely located atoms [95, 96]. Additionally, with the development of recent gridless sparse methods it was shown that the local coherence between nearly located atoms actually does not matter for the convex relaxation approach if the true DOAs are appropriately separated [97] (details will be provided in Section 6).

To improve the computational speed and accuracy, a heuristic grid refinement strategy was proposed that suggests using a coarse grid at the initial stage and then gradually refining it based on the latest estimates of the DOAs [67]. A grid selection approach was also proposed by quantifying the similarity between the atoms in a grid bin [93]. In particular, suppose that in (36) the DOA interval $(\theta_n - \frac{r}{2}, \theta_n + \frac{r}{2})$ is approximated by some grid point θ_n , where $r > 0$ denotes the grid interval. Then, on this interval the similarity is measured

by the rank of the matrix defined by

$$\mathbf{C}_n = \int_{\theta_n - \frac{r}{2}}^{\theta_n + \frac{r}{2}} \mathbf{a}(v) \mathbf{a}^H(v) dv. \quad (84)$$

If $\text{rank}(\mathbf{C}_n) \approx 1$, then it is said that the grid is dense enough; otherwise, a denser grid is required. However, a problem with this criterion is that it can only be evaluated heuristically.

In summary, grid selection is an important problem that affects the practical DOA estimation accuracy, the computational speed and the theoretical analysis. A completely satisfactory solution to this problem within the framework of the on-grid methods seems hard to obtain since there always exist mismatches between the adopted discrete grid points and the true continuous DOAs.

5 Off-Grid Sparse Methods

We have discussed the on-grid sparse methods in the preceding section, for which grid selection is a difficult problem and will inevitably result in grid mismatch. To resolve the grid mismatch problem, in this section we turn to the so-called off-grid sparse methods. In these methods, a grid is still required to perform sparse estimation but, unlike the on-grid methods, the DOA estimates are not restricted to be on the grid. We will mainly talk about two kinds of off-grid sparse methods: one is based on a fixed grid and joint estimation of the sparse signal and the grid offset, and the other relies on a dynamic grid. The main focus of the following discussions is on how to solve the grid mismatch.

5.1 Fixed Grid

5.1.1 Data Model

With a fixed grid $\bar{\theta} = \{\bar{\theta}_1, \dots, \bar{\theta}_{\bar{N}}\}$, an off-grid data model can be introduced as follows [98]. Suppose without loss of generality that $\bar{\theta}$ consists of uniformly spaced grid points with the grid interval $r = \theta_2 - \theta_1 \propto \frac{1}{\bar{N}}$. For any DOA θ_k , suppose $\bar{\theta}_{n_k}$ is the nearest grid point with $|\theta_k - \bar{\theta}_{n_k}| \leq \frac{r}{2}$. We approximate the steering vector/atom $\mathbf{a}(\theta_k)$ using a first-order Taylor expansion:

$$\mathbf{a}(\theta_k) \approx \mathbf{a}(\bar{\theta}_{n_k}) + \mathbf{b}(\bar{\theta}_{n_k}) (\theta_k - \bar{\theta}_{n_k}), \quad (85)$$

where $\mathbf{b}(\bar{\theta}_{n_k}) := \mathbf{a}'(\bar{\theta}_{n_k})$ (the derivative of $\mathbf{a}(\theta)$). Similar to (36), we then obtain the following data model:

$$\mathbf{Y} = \Phi(\beta) \mathbf{X} + \mathbf{E}, \quad (86)$$

where $\Phi(\beta) = \mathbf{A} + \mathbf{B} \text{diag}(\beta)$, $\mathbf{A} = [\mathbf{a}(\bar{\theta}_1), \dots, \mathbf{a}(\bar{\theta}_{\bar{N}})]$ is as defined previously, $\mathbf{B} = [\mathbf{b}(\bar{\theta}_1), \dots, \mathbf{b}(\bar{\theta}_{\bar{N}})]$ and $\beta = [\beta_1, \dots, \beta_{\bar{N}}] \in [-\frac{r}{2}, \frac{r}{2}]^{\bar{N}}$, with

$$x_n(t) = \begin{cases} s_k(t), & \text{if } \bar{\theta}_n = \bar{\theta}_{n_k}; \\ 0, & \text{otherwise,} \end{cases} \quad (87)$$

$$\beta_n = \begin{cases} \theta_k - \bar{\theta}_{n_k}, & \text{if } \bar{\theta}_n = \bar{\theta}_{n_k}; \\ 0, & \text{otherwise,} \end{cases} \quad n = 1, \dots, \bar{N}, \quad t = 1, \dots, L. \quad (88)$$

It follows from (86) that the DOA estimation problem can be formulated as sparse representation with uncertain parameters. In particular, once the row-sparse matrix \mathbf{X} and β can be estimated from \mathbf{Y} , then the DOAs can be estimated using the row-support of \mathbf{X} shifted by the offset β .

Compared to the on-grid model in (36), the additional grid offset parameters β_n , $n = 1, \dots, \bar{N}$ are introduced in the off-grid model in (86). Note that (86) reduces to (36) if $\beta = \mathbf{0}$. While (36) is based on a zeroth order approximation of the true data model, which causes grid mismatch, (86) can be viewed as a first-order approximation in which the grid mismatch can be partially compensated by jointly estimating the grid offset. Based on the off-grid model in (86), several methods have been proposed for DOA estimation by jointly estimating \mathbf{X} and β (see, e.g., [98–115]). Out of these methods we present the ℓ_1 -based optimization and SBL in the next subsections.

5.1.2 ℓ_1 Optimization

Inspired by the standard sparse signal recovery approach, several ℓ_1 optimization methods have been proposed to solve the off-grid DOA estimation problem. In [98], a sparse total least-squares (STLS) approach was proposed which, in the single snapshot case, solves the following LASSO-like problem:

$$\min_{\mathbf{x}, \beta} \lambda_1 \|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{y} - [\mathbf{A} + \mathbf{B} \text{diag}(\beta)] \mathbf{x}\|_2^2 + \lambda_2 \|\beta\|_2^2, \quad (89)$$

where λ_1 and λ_2 are regularization parameters. In (89), the prior information that $\beta \in [-\frac{\tau}{2}, \frac{\tau}{2}]^{\bar{N}}$ is not used. To heuristically control the magnitude of β , its power is also minimized. Note that the problem in (89) is nonconvex due to the bilinear term $\text{diag}(\beta) \mathbf{x}$. To solve (89), an alternating algorithm is adopted, iteratively solving for \mathbf{x} and β . Moreover, (89) can be easily extended to the multiple snapshot case by using $\ell_{2,1}$ optimization to exploit the joint sparsity in \mathbf{X} (as in the preceding section). A difficult problem of these methods is parameter tuning, i.e., how to choose λ_1 and λ_2 .

To exploit the prior knowledge that $\beta \in [-\frac{\tau}{2}, \frac{\tau}{2}]^{\bar{N}}$, the following BPDN-like formulation was proposed in the single snapshot case [100]:

$$\min_{\mathbf{x}, \beta \in [-\frac{\tau}{2}, \frac{\tau}{2}]^{\bar{N}}} \|\mathbf{x}\|_1, \text{ subject to } \|\mathbf{y} - [\mathbf{A} + \mathbf{B} \text{diag}(\beta)] \mathbf{x}\|_2 \leq \eta. \quad (90)$$

Note that (90) can be easily extended to the multiple snapshot case by using the $\ell_{2,1}$ norm. In (90), η can be set according to information about the noise level and a possible estimate of the modeling error. Similar to (89), (90) is nonconvex, and a similar alternating algorithm can be implemented to monotonically decrease the value of the objective function. Note that if β is initialized as a zero vector, then the first iteration coincides with the standard BPDN.

It was shown in [100] that if the matrix $[\mathbf{A}, \mathbf{B}]$ satisfies a certain RIP condition, then both \mathbf{x} and β can be stably reconstructed, as in the standard sparse representation problem, with the reconstruction error being proportional to the noise level η . This means that in the ideal case of $\eta = 0$ (assuming there is no noise or modeling error), \mathbf{x} and β can be exactly recovered. A key step in showing this result is reformulating (90) as

$$\min_{\mathbf{x}, \beta \in [-\frac{\tau}{2}, \frac{\tau}{2}]^{\bar{N}}} \|\mathbf{x}\|_1, \text{ subject to } \left\| \mathbf{y} - [\mathbf{A} \quad \mathbf{B}] \begin{bmatrix} \mathbf{x} \\ \beta \odot \mathbf{x} \end{bmatrix} \right\|_2 \leq \eta, \quad (91)$$

where \odot denotes the element-wise product. Although the RIP condition cannot be easily applied to the case of dense grid, the aforementioned result implies, to some extent, the superior performance of this off-grid optimization method as compared to the on-grid approach.

Following the lead of [100], a convex optimization method was proposed in [102] by exploiting the joint sparsity of \mathbf{x} and $\mathbf{v} := \beta \odot \mathbf{x}$. In particular, the following problem was formulated:

$$\min_{\mathbf{x}, \mathbf{v}} \lambda \|\begin{bmatrix} \mathbf{x} & \mathbf{v} \end{bmatrix}\|_{2,1} + \frac{1}{2} \left\| \mathbf{y} - [\mathbf{A} \quad \mathbf{B}] \begin{bmatrix} \mathbf{x} \\ \mathbf{v} \end{bmatrix} \right\|_2^2, \quad (92)$$

which is equivalent to the following problem, for appropriate parameter choices:

$$\min_{\mathbf{x}, \mathbf{v}} \|\begin{bmatrix} \mathbf{x} & \mathbf{v} \end{bmatrix}\|_{2,1}, \text{ subject to } \left\| \mathbf{y} - \begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{v} \end{bmatrix} \right\|_2 \leq \eta. \quad (93)$$

This approach is advantageous in that it is convex and can be globally solved in a polynomial time, with similar theoretical guarantees as provided in [100]. However, it is worth noting that the prior knowledge on β cannot be exploited in this method. Additionally, the obtained solution for $\beta_n = \frac{v_n}{x_n}$ might not even be real. To resolve this problem, [102] suggests a two-stage solution: 1) obtain \mathbf{x} from (92), and 2) fix \mathbf{x} and solve for β by minimizing $\|\mathbf{y} - [\mathbf{A} + \mathbf{B}\text{diag}(\beta)]\mathbf{x}\|_2$.

5.1.3 Sparse Bayesian Learning

A systematic approach to off-grid DOA estimation, called off-grid sparse Bayesian inference (OGSBI), was proposed in [101] within the framework of SBL in the multiple snapshot case. In order to estimate the additional parameter β , it is assumed that $\beta_n, n = 1, \dots, \bar{N}$ are i.i.d. uniformly distributed on the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$. In the resulting EM algorithm, the posterior distribution of the row-sparse signal \mathbf{X} can be computed in the expectation step as in the standard SBL. In the maximization step, β is also updated, in addition to updating the power \mathbf{p} of the row-sparse signal and the noise variance σ . As in the standard SBL, the likelihood is guaranteed to monotonically increase and hence convergence of the algorithm can be obtained.

5.2 Dynamic Grid

5.2.1 Data Model

The data model now uses a dynamic grid $\bar{\boldsymbol{\theta}}$ in the sense that the grid points $\theta_n, n = 1, \dots, \bar{N}$ are not fixed:

$$\mathbf{Y} = \mathbf{A}(\bar{\boldsymbol{\theta}})\mathbf{X} + \mathbf{E}. \quad (94)$$

For this model we need to jointly estimate the row-sparse matrix \mathbf{X} and the grid $\bar{\boldsymbol{\theta}}$. Once they are obtained, the DOAs are estimated using those grid points of $\bar{\boldsymbol{\theta}}$ corresponding to the nonzero rows of \mathbf{X} . Since $\bar{\theta}_n$'s are estimated from the data and can be any values in the continuous DOA domain, this off-grid data model is accurate and does not suffer from grid mismatch. However, the difficulty lies in designing an algorithm for the joint estimation of \mathbf{X} and $\bar{\boldsymbol{\theta}}$, due to the nonlinearity in $\bar{\boldsymbol{\theta}}$. Note that the following algorithms that we will introduce are designated as off-grid methods, instead of gridless, since grid selection remains involved in them (e.g., choice of \bar{N} and initialization of $\bar{\boldsymbol{\theta}}$), which affects the computational speed and accuracy of the algorithms.

5.2.2 Algorithms

Several algorithms have been proposed based on the data model in (94). The first class is within the framework of SBL (see, e.g., [116–119]). But instead of using the EM algorithm as previously, a variational EM algorithm (or variational Bayesian inference) is typically exploited to carry out the sparse signal and parameter estimation. The reason is that the posterior distribution of the sparse vector \mathbf{x} usually cannot be explicitly computed here, and that distribution is required by the EM but not by the variational EM. The main difficulty of these algorithms is the update of $\boldsymbol{\theta}$ due to the strong nonlinearity. Because closed-form solutions are not available, only numerical approaches can be used.

Another class of methods uses ℓ_1 optimization. In the single snapshot case, as an example, the paper [94] used a small $\overline{N} \geq K$ and attempted to solve the following ℓ_1 optimization problem by iteratively updating \mathbf{x} and $\boldsymbol{\theta}$:

$$\min_{\mathbf{x}, \boldsymbol{\theta}} \lambda \|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta}) \mathbf{x}\|_2^2. \quad (95)$$

To avoid the possible convergence of some $\bar{\theta}_n$'s to the same value, an additional (nonconvex) term $g(\bar{\boldsymbol{\theta}})$ is included to penalize closely located parameters:

$$\min_{\mathbf{x}, \boldsymbol{\theta}} \lambda_1 \|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{y} - \mathbf{A}(\bar{\boldsymbol{\theta}}) \mathbf{x}\|_2^2 + \lambda_2 g(\bar{\boldsymbol{\theta}}), \quad (96)$$

where λ_1 and λ_2 are regularization parameters that need be tuned. Note that both (95) and (96) are nonconvex. Even for given \mathbf{x} , it is difficult to solve for $\bar{\boldsymbol{\theta}}$. Moreover, parameter tuning is tricky. Note that ℓ_q , $q < 1$ optimization was also considered in [94] to enhance sparsity but it suffers from similar problems.

To promote sparsity, similar to ℓ_1 optimization, the following problem was proposed in [120, 121]:

$$\min_{\mathbf{x}, \boldsymbol{\theta}} \sum_{n=1}^{\overline{N}} \lambda \log(|x_n|^2 + \epsilon) + \|\mathbf{y} - \mathbf{A}(\bar{\boldsymbol{\theta}}) \mathbf{x}\|_2^2. \quad (97)$$

To locally solve (97), \mathbf{x} and $\bar{\boldsymbol{\theta}}$ are iteratively updated. To solve for \mathbf{x} in closed form, the first term of the objective in (97) is replaced by a quadratic surrogate function that guarantees the decrease of the objective. The gradient descent method is then used to solve for $\bar{\boldsymbol{\theta}}$. While it is generally difficult to choose λ , [121] suggested setting λ proportional to the inverse of the noise variance, leaving a constant coefficient to be tuned.

To conclude, in this section we introduced several off-grid sparse optimization methods for DOA estimation. By imposing fewer constraints on the DOA candidates, the grid mismatch encountered in the on-grid sparse methods can be overcome. However, this introduces more variables that need be estimated and complicates the algorithm design. As a consequence, most of the presented algorithms involve nonconvex optimization and thus only local convergence can be guaranteed (except for the algorithm in [102]). Moreover, few theoretical guarantees can be obtained for most of the algorithms (see, however, [100, 102]).

6 Gridless Sparse Methods

In this section, we present several recent DOA estimation approaches that are designated as the gridless sparse methods. As their name suggests, these methods do not require gridding of the direction domain. Instead, they directly operate in the continuous domain and therefore can completely resolve the grid mismatch problem. Moreover, they are convex and have strong theoretical guarantees. However, so far, this kind of methods can only be applied to uniform or sparse linear arrays. Therefore, naturally, in this section we treat the DOA estimation problem as frequency estimation, following the discussions in Section 2.

The rest of this section is organized as follows. We first revisit the data model in the context of ULAs and SLAs. We then introduce a mathematical tool known as the Vandermonde decomposition of Toeplitz covariance matrices, which is crucial for most gridless sparse methods. Finally, we discuss a number of gridless sparse methods for DOA/frequency estimation in the case of a single snapshot, followed by the case of multiple snapshots. The atomic norm and gridless SPICE methods will be particularly highlighted.

6.1 Data Model

For convenience, we restate the data model that will be used in this section. For an M -element ULA, the array data are modeled as:

$$\mathbf{Y} = \mathbf{A}(\mathbf{f})\mathbf{S} + \mathbf{E}, \quad (98)$$

where $f_k \in \mathbb{T}$, $k = 1, \dots, K$ are the frequencies of interest, which have a one-to-one relationship to the DOAs, $\mathbf{A}(\mathbf{f}) = [\mathbf{a}(f_1), \dots, \mathbf{a}(f_K)] \in \mathbb{C}^{M \times K}$ is the array manifold matrix, and $\mathbf{a}(f) = [1, e^{i2\pi f}, \dots, e^{i2\pi(M-1)f}]^T \in \mathbb{C}^M$ is the steering vector.

For an M -element SLA, suppose that the array is obtained from an N -element virtual ULA by retaining the antennas indexed by the set $\Omega = \{\Omega_1, \dots, \Omega_M\}$, where $N \geq M$ and $1 \leq \Omega_1 < \dots < \Omega_M \leq N$. In this case, we can view (98) as the data model with the virtual ULA. Then the data model of the SLA is given by

$$\mathbf{Y}_\Omega = \mathbf{A}_\Omega(\mathbf{f})\mathbf{S} + \mathbf{E}_\Omega. \quad (99)$$

Therefore, (98) can be considered as a special case of (99) in which $M = N$ and $\Omega = \{1, \dots, N\}$. Given \mathbf{Y} (or \mathbf{Y}_Ω and Ω), the objective is to estimate the frequencies f_k 's (note that the source number K is usually unknown).

In the single snapshot case the above problem coincides with the line spectral estimation problem. Since the first gridless sparse methods were developed for the latter problem, we present them in the single snapshot case and then discuss how they can be extended to the multiple snapshot case by exploiting the joint sparsity of the snapshots. Before doing that, an important mathematical tool is introduced in the following subsection.

6.2 Vandermonde Decomposition of Toeplitz Covariance Matrices

The Vandermonde decomposition of Toeplitz covariance matrices plays an important role in this section. This classical result was discovered by Carathéodory and Fejér in 1911 [122]. It has become important in the area of data analysis and signal processing since the 1970s when it was rediscovered by Pisarenko and used for frequency retrieval from the data covariance matrix [2]. From then on, the Vandermonde decomposition has formed the basis of a prominent subset of methods for frequency and DOA estimation, viz. the subspace-based methods. To see why it is so, let us consider the data model in (98) and assume uncorrelated sources. In the noiseless case, the data covariance matrix is given by

$$\mathbf{R} = \mathbb{E}\mathbf{y}(t)\mathbf{y}^H(t) = \mathbf{A}(\mathbf{f}) \text{diag}(\mathbf{p}) \mathbf{A}^H(\mathbf{f}), \quad (100)$$

where $p_k > 0$, $k = 1, \dots, K$ are the powers of the sources. It can be easily verified that \mathbf{R} is a (Hermitian) Toeplitz matrix that can be written as:

$$\mathbf{R} = \mathbf{T}(\mathbf{u}) := \begin{bmatrix} u_1 & u_2 & \cdots & u_N \\ u_2^* & u_1 & \cdots & u_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ u_N^* & u_{N-1}^* & \cdots & u_1 \end{bmatrix}, \quad (101)$$

where $\mathbf{u} \in \mathbb{C}^N$. Moreover, \mathbf{R} is PSD and has rank K under the assumption that $K < N$. The Vandermonde decomposition result states that any rank-deficient, PSD Toeplitz matrix \mathbf{T} can be uniquely decomposed as in (100). Equivalently stated, this means that the frequencies can be exactly retrieved from the data covariance matrix. Formally, the result is stated in the following theorem, a proof of which (inspired by [123]) is also provided; note that the proof suggests a way of computing the decomposition.

Theorem 6.1. Any PSD Toeplitz matrix $\mathbf{T}(\mathbf{u}) \in \mathbb{C}^{N \times N}$ of rank $r \leq N$ admits the following r -atomic Vandermonde decomposition:

$$\mathbf{T} = \sum_{k=1}^r p_k \mathbf{a}(f_k) \mathbf{a}^H(f_k) = \mathbf{A}(\mathbf{f}) \text{diag}(\mathbf{p}) \mathbf{A}^H(\mathbf{f}), \quad (102)$$

where $p_k > 0$, and $f_k \in \mathbb{T}$, $k = 1, \dots, r$ are distinct. Moreover, the decomposition in (102) is unique if $r < N$.

Proof. We first consider the case of $r = \text{rank}(\mathbf{T}) \leq N - 1$. Since $\mathbf{T} \geq \mathbf{0}$, there exists $\mathbf{V} \in \mathbb{C}^{N \times r}$ satisfying $\mathbf{T} = \mathbf{V}\mathbf{V}^H$. Let \mathbf{V}_{-N} and \mathbf{V}_{-1} be the matrices obtained from \mathbf{V} by removing its last and first row, respectively. By the structure of \mathbf{T} , we have that $\mathbf{V}_{-N}\mathbf{V}_{-N}^H = \mathbf{V}_{-1}\mathbf{V}_{-1}^H$. Thus there must exist an $r \times r$ unitary matrix \mathbf{Q} satisfying $\mathbf{V}_{-1} = \mathbf{V}_{-N}\mathbf{Q}$ (see, e.g., [124, Theorem 7.3.11]). It follows that $\mathbf{V}_j = \mathbf{V}_1\mathbf{Q}^{j-1}$, $j = 2, \dots, N$ and therefore,

$$u_j = \mathbf{V}_1\mathbf{Q}^{1-j}\mathbf{V}_1^H, \quad j = 1, \dots, N. \quad (103)$$

Next, write the eigen-decomposition of the unitary matrix \mathbf{Q} , which is guaranteed to exist, as

$$\mathbf{Q} = \tilde{\mathbf{Q}}\text{diag}(z_1, \dots, z_r)\tilde{\mathbf{Q}}^H, \quad (104)$$

where $\tilde{\mathbf{Q}}$ is also an $r \times r$ unitary matrix and z_k 's are the eigenvalues. Since the eigenvalues of a unitary matrix must have unit magnitude, we can find $f_k \in \mathbb{T}$, $k = 1, \dots, r$ satisfying $z_k = e^{i2\pi f_k}$, $k = 1, \dots, r$. Inserting (104) into (103) and letting $p_k = \left| \mathbf{V}_1\tilde{\mathbf{Q}}_{:,k} \right|^2 > 0$, $k = 1, \dots, r$, where $\tilde{\mathbf{Q}}_{:,k}$ denotes the k th column of $\tilde{\mathbf{Q}}$, we have that

$$u_j = \sum_{k=1}^r p_k e^{-i2\pi(j-1)f_k}. \quad (105)$$

It follows that (102) holds. Moreover, f_k , $k = 1, \dots, r$ are distinct since otherwise $\text{rank}(\mathbf{T}) < r$, which cannot be true.

We now consider the case of $r = N$ for which $\mathbf{T} > \mathbf{0}$. Let us arbitrarily choose $f_N \in \mathbb{T}$ and let $p_N = (\mathbf{a}^H(f_N)\mathbf{T}^{-1}\mathbf{a}(f_N))^{-1} > 0$. Moreover, we define a new vector $\mathbf{u}' \in \mathbb{C}^N$ by

$$u'_j = u_j - p_N e^{-i2\pi(j-1)f_N}. \quad (106)$$

It can be readily verified that

$$\mathbf{T}(\mathbf{u}') = \mathbf{T}(\mathbf{u}) - p_N \mathbf{a}(f_N) \mathbf{a}^H(f_N), \quad (107)$$

$$\mathbf{T}(\mathbf{u}') \geq \mathbf{0}, \quad (108)$$

$$\text{rank}(\mathbf{T}(\mathbf{u}')) = N - 1. \quad (109)$$

Therefore, from the result in the case of $r \leq N - 1$ proven above, $\mathbf{T}(\mathbf{u}')$ admits a Vandermonde decomposition as in (102) with $r = N - 1$. It then follows from (107) that $\mathbf{T}(\mathbf{u})$ admits a Vandermonde decomposition with N "atoms".

We finally show the uniqueness in the case of $r \leq N - 1$. To do so, suppose there exists another decomposition $\mathbf{T} = \mathbf{A}(\mathbf{f}')\mathbf{P}'\mathbf{A}^H(\mathbf{f}')$ in which $p'_j > 0$, $j = 1, \dots, r$ and $f'_j \in \mathbb{T}$ are distinct. It follows from the equation

$$\mathbf{A}(\mathbf{f}')\mathbf{P}'\mathbf{A}^H(\mathbf{f}') = \mathbf{A}(\mathbf{f})\mathbf{P}\mathbf{A}^H(\mathbf{f}) \quad (110)$$

that there exists an $r \times r$ unitary matrix \mathbf{Q}' such that $\mathbf{A}(\mathbf{f}') \mathbf{P}'^{\frac{1}{2}} = \mathbf{A}(\mathbf{f}) \mathbf{P}^{\frac{1}{2}} \mathbf{Q}'$ and therefore,

$$\mathbf{A}(\mathbf{f}') = \mathbf{A}(\mathbf{f}) \mathbf{P}^{\frac{1}{2}} \mathbf{Q}' \mathbf{P}'^{-\frac{1}{2}}. \quad (111)$$

This means that for every $j = 1, \dots, r$, $\mathbf{a}(f'_j)$ lies in the range space spanned by $\{\mathbf{a}(f_k)\}_{k=1}^r$. By the fact that $r \leq N - 1$ and that any N atoms $\mathbf{a}(f_k)$ with distinct f_k 's are linearly independent, we have that $f'_j \in \{f_k\}_{k=1}^r$ and thus the two sets $\{f'_j\}_{j=1}^r$ and $\{f_k\}_{k=1}^r$ are identical. It follows that the above two decompositions of $\mathbf{T}(\mathbf{u})$ must be identical. ■

Note that the proof of Theorem 6.1 provides a computational approach to the Vandermonde decomposition. We simply consider the case of $r \leq N - 1$, since in the case of $r = N$ we can arbitrarily choose $f_N \in \mathbb{T}$ first. We use the following result:

$$(\mathbf{V}_{-N}^H \mathbf{V}_{-1} - z_k \mathbf{V}_{-N}^H \mathbf{V}_{-N}) \tilde{\mathbf{Q}}_{:k} = 0, \quad (112)$$

which can be shown along the lines of the above proof. To retrieve the frequencies and the powers from \mathbf{T} , we first compute $\mathbf{V} \in \mathbb{C}^{N \times r}$ satisfying $\mathbf{T} = \mathbf{V} \mathbf{V}^H$ using, e.g., the Cholesky decomposition. After that, we use (112) and compute z_k and $\tilde{\mathbf{Q}}_{:k}$, $k = 1, \dots, r$ as the eigenvalues and the normalized eigenvectors of the matrix pencil $(\mathbf{V}_{-N}^H \mathbf{V}_{-1}, \mathbf{V}_{-N}^H \mathbf{V}_{-N})$. Finally, we obtain $f_k = \frac{1}{2\pi} \Im \ln z_k \in \mathbb{T}$ and $p_k = \|\mathbf{V}_1 \tilde{\mathbf{Q}}_{:k}\|^2$, $k = 1, \dots, r$, where \Im gives the imaginary part of its argument and \mathbf{V}_1 is the first row of \mathbf{V} . In fact, this matrix pencil approach is similar to the ESPRIT algorithm that computes the frequency estimates from an estimate of the data covariance matrix.

In the presence of homoscedastic noise, the data covariance matrix \mathbf{R} remains Toeplitz. In this case, it is natural to decompose the Toeplitz covariance matrix as the sum of the signal covariance and the noise covariance. Consequently, the following corollary of Theorem 6.1 can be useful in such a case. The proof is straightforward and will be omitted.

Corollary 6.1. Any PSD Toeplitz matrix $\mathbf{T}(\mathbf{u}) \in \mathbb{C}^{N \times N}$ can be uniquely decomposed as:

$$\mathbf{T} = \sum_{k=1}^r p_k \mathbf{a}(f_k) \mathbf{a}^H(f_k) + \sigma \mathbf{I} = \mathbf{A}(\mathbf{f}) \text{diag}(\mathbf{p}) \mathbf{A}^H(\mathbf{f}) + \sigma \mathbf{I}, \quad (113)$$

where $\sigma = \lambda_{\min}(\mathbf{T})$ (the smallest eigenvalue of \mathbf{T}), $r = \text{rank}(\mathbf{T} - \sigma \mathbf{I}) < N$, $p_k > 0$, and $f_k \in \mathbb{T}$, $k = 1, \dots, r$ are distinct.

Remark 6.1. Note that the uniqueness of the decomposition in Corollary 6.1 is guaranteed by the condition that $\sigma = \lambda_{\min}(\mathbf{T})$. If the condition is violated by letting $0 \leq \sigma < \lambda_{\min}(\mathbf{T})$ (in such a case \mathbf{T} has full rank and $r \geq N$), then the decomposition in (113) cannot be unique.

The Vandermonde decomposition of Toeplitz covariance matrices forms an important tool in several recently proposed gridless sparse methods. In particular, these methods transform the frequency estimation problem into the estimation of a PSD Toeplitz matrix in which the frequencies are encoded. Once the matrix is computed, the frequencies can be retrieved from its Vandermonde decomposition. Therefore, these gridless sparse methods can be viewed as being covariance-based by interpreting the Toeplitz matrix as the data covariance matrix (though it might not be, since certain statistical assumptions may not be satisfied). In contrast to conventional subspace methods that estimate the frequencies directly from the sample covariance matrix, the gridless methods utilize more sophisticated optimization approaches to estimate the data covariance matrix by exploiting its special structures, e.g., Toeplitz, low rank and PSDness, and therefore are expected to achieve superior performance.

6.3 The Single Snapshot Case

In this subsection we introduce several gridless sparse methods for DOA/frequency estimation in the single snapshot case (a.k.a. the line spectral estimation problem). Two kinds of methods will be discussed: deterministic optimization methods, e.g., the atomic norm and the Hankel-based nuclear norm methods [81, 97, 125–135], and a covariance fitting method that is a gridless version of SPICE [82, 130, 136–140]. The connections between these methods will also be investigated. By ‘deterministic’ we mean that we do not make any statistical assumptions on the signal of interest. Instead, the signal is deterministic and it is sought as the sparsest candidate, measured by a certain sparse metric, among a prescribed set.

6.3.1 A General Framework for Deterministic Methods

In the single snapshot case, corresponding to (99), the data model in the SLA case is given by:

$$\mathbf{y}_\Omega = \mathbf{z}_\Omega + \mathbf{e}_\Omega, \quad \mathbf{z} := \mathbf{A}(\mathbf{f}) \mathbf{s}, \quad (114)$$

where \mathbf{z} denotes the noiseless signal. Note that the ULA is a special case with $\Omega = \{1, \dots, N\}$. For deterministic sparse methods, in general, we need to solve a constrained optimization problem of the following form:

$$\min_{\mathbf{z}} \mathcal{M}(\mathbf{z}), \quad \text{subject to } \|\mathbf{z}_\Omega - \mathbf{y}_\Omega\|_2 \leq \eta, \quad (115)$$

where the noise is assumed to be bounded: $\|\mathbf{e}_\Omega\|_2 \leq \eta$. Alternatively, we solve a regularized optimization problem given by:

$$\min_{\mathbf{z}} \lambda \mathcal{M}(\mathbf{z}) + \frac{1}{2} \|\mathbf{z}_\Omega - \mathbf{y}_\Omega\|_2^2, \quad (116)$$

where the noise is typically assumed to be Gaussian and $\lambda > 0$ is a regularization parameter. In the extreme noiseless case, by letting $\eta \rightarrow 0$ and $\lambda \rightarrow 0$, both (115) and (116) reduce to the following problem:

$$\min_{\mathbf{z}} \mathcal{M}(\mathbf{z}), \quad \text{subject to } \mathbf{z}_\Omega = \mathbf{y}_\Omega. \quad (117)$$

In (115)-(117), \mathbf{z} is the sinusoidal signal of interest. Furthermore, $\mathcal{M}(\mathbf{z})$ denotes a sparse metric that is defined such that by minimizing $\mathcal{M}(\mathbf{z})$ the number of components/atoms $\mathbf{a}(\mathbf{f})$ used to express \mathbf{z} is reduced, and these atoms give the frequency estimates. We next discuss different choices of $\mathcal{M}(\mathbf{z})$.

6.3.2 Atomic ℓ_0 Norm

To promote sparsity to the greatest extent possible, inspired by the literature on sparse recovery and compressed sensing, the natural choice of $\mathcal{M}(\mathbf{z})$ is an ℓ_0 norm like sparse metric, referred to as the atomic ℓ_0 (pseudo-)norm. Let us formally define the set of atoms used here:

$$\mathcal{A} := \{\mathbf{a}(f, \phi) := \mathbf{a}(f) \phi : f \in \mathbb{T}, \phi \in \mathbb{C}, |\phi| = 1\}. \quad (118)$$

It is evident from (114) that the true signal \mathbf{z} is a linear combination of K atoms in the atomic set \mathcal{A} . The atomic ℓ_0 (pseudo-)norm, denoted by $\|\mathbf{z}\|_{\mathcal{A},0}$, is defined as the minimum number of atoms in \mathcal{A} that can synthesize \mathbf{z} :

$$\begin{aligned} \|\mathbf{z}\|_{\mathcal{A},0} &:= \inf_{c_k, f_k, \phi_k} \left\{ \mathcal{K} : \mathbf{z} = \sum_{k=1}^{\mathcal{K}} \mathbf{a}(f_k, \phi_k) c_k, f_k \in \mathbb{T}, |\phi_k| = 1, c_k > 0 \right\} \\ &= \inf_{f_k, s_k} \left\{ \mathcal{K} : \mathbf{z} = \sum_{k=1}^{\mathcal{K}} \mathbf{a}(f_k) s_k, f_k \in \mathbb{T} \right\}. \end{aligned} \quad (119)$$

To provide a finite-dimensional formulation for $\|z\|_{\mathcal{A},0}$, the Vandermonde decomposition of Toeplitz covariance matrices is invoked. To be specific, let \mathbf{T} be a Toeplitz matrix and impose the condition that

$$\begin{bmatrix} x & z^H \\ z & \mathbf{T} \end{bmatrix} \geq \mathbf{0}, \quad (120)$$

where x is a free variable to be optimized. It follows from (120) that \mathbf{T} is PSD and thus admits a rank (\mathbf{T})-atomic Vandermonde decomposition. Moreover, z lies in the range space of \mathbf{T} . Therefore, z can be expressed by rank (\mathbf{T}) atoms. This means that the atomic ℓ_0 norm is linked to the rank of \mathbf{T} . Formally, we have the following result.

Theorem 6.2 ([127]). $\|z\|_{\mathcal{A},0}$ defined in (119) equals the optimal value of the following rank minimization problem:

$$\min_{x, \mathbf{u}} \text{rank}(\mathbf{T}(\mathbf{u})), \text{ subject to (120)}. \quad (121)$$

By Theorem 6.2, the atomic ℓ_0 norm method needs to solve a rank minimization problem that, as might have been expected, cannot be easily solved. By the rank minimization formulation, the frequencies of interest are actually encoded in the PSD Toeplitz matrix $\mathbf{T}(\mathbf{u})$. If $\mathbf{T}(\mathbf{u})$ can be solved for, then the frequencies can be retrieved from its Vandermonde decomposition. Therefore, the Toeplitz matrix $\mathbf{T}(\mathbf{u})$ in (121) can be viewed as the covariance matrix of the noiseless signal z as if certain statistical assumptions were satisfied (however, those assumptions are not required here). Note that the Toeplitz structure of the covariance matrix is explicitly enforced, the PSDness is imposed by the constraint in (120), and the low-rankness is the objective.

6.3.3 Atomic Norm

A practical choice of the sparse metric $\mathcal{M}(z)$ is the atomic norm that is a convex relaxation of the atomic ℓ_0 norm. The resulting optimization problems in (115)-(117) are referred to as atomic norm minimization (ANM). The concept of atomic norm was first proposed in [141] and it generalizes several norms commonly used for sparse representation and recovery, e.g., the ℓ_1 norm and the nuclear norm, for appropriately chosen atoms. The atomic norm is basically equivalent to the total variation norm [142] that was adopted, e.g., in [97]. We have decided to use the atomic norm in this article since it is simpler to present and easier to understand. Formally, the atomic norm is defined as the gauge function of $\text{conv}(\mathcal{A})$, the convex hull of \mathcal{A} [141]:

$$\begin{aligned} \|z\|_{\mathcal{A}} &:= \inf \{t > 0 : z \in t\text{conv}(\mathcal{A})\} \\ &= \inf_{c_k, f_k, \phi_k} \left\{ \sum_k c_k : z = \sum_k \mathbf{a}(f_k, \phi_k) c_k, f_k \in \mathbb{T}, |\phi_k| = 1, c_k > 0 \right\} \\ &= \inf_{f_k, s_k} \left\{ \sum_k |s_k| : z = \sum_k \mathbf{a}(f_k) s_k, f_k \in \mathbb{T} \right\}. \end{aligned} \quad (122)$$

By definition, the atomic norm can be viewed as a continuous counterpart of the ℓ_1 norm used in the discrete setting. Different from the ℓ_1 norm, however, it is unclear how to compute the atomic norm from the definition. In fact, initially this has been a major obstacle in applying the atomic norm technique [141, 143]. To solve this problem, a computationally efficient SDP formulation of $\|z\|_{\mathcal{A}}$ is provided in the following result. A proof of the result is also provided, which helps illustrate how the frequencies can be obtained.

Theorem 6.3 ([127]). $\|z\|_{\mathcal{A}}$ defined in (122) equals the optimal value of the following SDP:

$$\min_{x, \mathbf{u}} \frac{1}{2}x + \frac{1}{2}u_1, \text{ subject to (120).} \quad (123)$$

Proof. Let F be the optimal value of the objective in (123). We need to show that $F = \|z\|_{\mathcal{A}}$.

We first show that $F \leq \|z\|_{\mathcal{A}}$. To do so, let $z = \sum_k c_k \mathbf{a}(f_k, \phi_k) = \sum_k \mathbf{a}(f_k) s_k$ be an atomic decomposition of z . Then let \mathbf{u} be such that $\mathbf{T}(\mathbf{u}) = \sum_k c_k \mathbf{a}(f_k) \mathbf{a}^H(f_k)$ and $x = \sum_k c_k$. It follows that

$$\begin{bmatrix} x & z^H \\ z & \mathbf{T} \end{bmatrix} = \sum_k c_k \begin{bmatrix} \phi_k^* \\ \mathbf{a}(f_k) \end{bmatrix} \begin{bmatrix} \phi_k^* \\ \mathbf{a}(f_k) \end{bmatrix}^H \geq \mathbf{0}. \quad (124)$$

Therefore, x and \mathbf{u} constructed as above form a feasible solution to the problem in (123), at which the objective value equals

$$\frac{1}{2}x + \frac{1}{2}u_1 = \sum_k c_k. \quad (125)$$

It follows that $F \leq \sum_k c_k$. Since the inequality holds for any atomic decomposition of z , we have that $F \leq \|z\|_{\mathcal{A}}$ by the definition of the atomic norm.

On the other hand, suppose that $(\hat{x}, \hat{\mathbf{u}})$ is an optimal solution to the problem in (123). By the fact that $\mathbf{T}(\hat{\mathbf{u}}) \geq \mathbf{0}$ and applying Theorem 6.1, we have that $\mathbf{T}(\hat{\mathbf{u}})$ admits a Vandermonde decomposition as in (102) with (r, p_k, f_k) denoted by $(\hat{r}, \hat{p}_k, \hat{f}_k)$. Moreover, since $\begin{bmatrix} \hat{x} & z^H \\ z & \mathbf{T}(\hat{\mathbf{u}}) \end{bmatrix} \geq \mathbf{0}$, we have that z lies in the range space of $\mathbf{T}(\hat{\mathbf{u}})$ and thus has the following atomic decomposition:

$$z = \sum_{k=1}^{\hat{r}} \hat{c}_k \mathbf{a}(\hat{f}_k, \hat{\phi}_k) = \sum_{k=1}^{\hat{r}} \mathbf{a}(\hat{f}_k) \hat{s}_k. \quad (126)$$

Moreover, it holds that

$$\hat{x} \geq z^H [\mathbf{T}(\hat{\mathbf{u}})]^\dagger z = \sum_{k=1}^{\hat{r}} \frac{\hat{c}_k^2}{\hat{p}_k}, \quad (127)$$

$$\hat{u}_0 = \sum_{k=1}^{\hat{r}} \hat{p}_k. \quad (128)$$

It therefore follows that

$$\begin{aligned} F &= \frac{1}{2}\hat{x} + \frac{1}{2}\hat{u}_1 \\ &\geq \frac{1}{2} \sum_k \frac{\hat{c}_k^2}{\hat{p}_k} + \frac{1}{2} \sum_k \hat{p}_k \\ &\geq \sum_k \hat{c}_k \\ &\geq \|z\|_{\mathcal{A}}. \end{aligned} \quad (129)$$

Combining (129) and the inequality $F \leq \|z\|_{\mathcal{A}}$ that was shown previously, we conclude that $F = \|z\|_{\mathcal{A}}$, which completes the proof. It is worth noting that by (129) we must have that $\hat{p}_k = \hat{c}_k = |\hat{s}_k|$ and $\|z\|_{\mathcal{A}} = \sum_k \hat{c}_k = \sum_k |\hat{s}_k|$. Therefore, the atomic decomposition in (126) achieves the atomic norm. ■

Interestingly (but not surprisingly), the SDP in (123) is actually a convex relaxation of the rank minimization problem in (121). Concretely, the second term $\frac{1}{2}u_1$ in the objective function in (123) is actually the nuclear norm or the trace norm of $\mathbf{T}(\mathbf{u})$ (up to a scaling factor), which is a commonly used convex relaxation of the matrix rank, while the first term $\frac{1}{2}x$ is a regularization term that prevents a trivial solution.

Similar to the atomic ℓ_0 norm, the frequencies in the atomic norm approach are also encoded in the Toeplitz matrix $\mathbf{T}(\mathbf{u})$. Once the resulting SDP problem is solved, the frequencies can be retrieved from the Vandermonde decomposition of $\mathbf{T}(\mathbf{u})$. Therefore, similar to the ℓ_0 norm, the atomic norm can also be viewed as being covariance-based. The only difference lies in enforcing the low-rankness of the ‘covariance’ matrix $\mathbf{T}(\mathbf{u})$. The atomic ℓ_0 norm directly uses the rank function that exploits the low-rankness to the greatest extent possible but cannot be practically solved. In contrast to this, the atomic norm uses a convex relaxation, the nuclear norm (or the trace norm), and provides a practically feasible approach.

In the absence of noise, the theoretical performance of ANM has been studied in [97, 127]. In the case of ULA where all the entries of \mathbf{y} are observed, the ANM problem derived from (117) actually admits a trivial solution $\mathbf{z} = \mathbf{y}$. But the following SDP resulting from (123) still makes sense and can be used for frequency estimation:

$$\min_{x, \mathbf{u}} \frac{1}{2}x + \frac{1}{2}u_1, \text{ subject to } \begin{bmatrix} x & \mathbf{y}^H \\ \mathbf{y} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (130)$$

Let $\mathcal{T} := \{f_1, \dots, f_K\}$ and define the minimum separation of \mathcal{T} as the closest wrap-around distance between any two elements:

$$\Delta_{\mathcal{T}} = \inf_{1 \leq j \neq k \leq K} \min \{|f_j - f_k|, 1 - |f_j - f_k|\}. \quad (131)$$

The following theoretical guarantee for the atomic norm is provided in [97].

Theorem 6.4 ([97]). $\mathbf{y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$ is the unique atomic decomposition satisfying $\|\mathbf{y}\|_{\mathcal{A}} = \sum_{j=1}^K c_j$ if $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$ and $N \geq 257$.

By Theorem 6.4, in the noiseless case the frequencies can be exactly recovered by solving the SDP in (130) if the frequencies are separated by at least $\frac{4}{N}$ (note that this frequency separation condition has recently been relaxed to $\frac{2.52}{N}$ in [133]). Moreover, the condition $N \geq 257$ is a technical requirement that should not pose any serious problem in practice.

In the SLA case, the SDP resulting from (117) is given by:

$$\min_{x, \mathbf{u}, \mathbf{z}} \frac{1}{2}x + \frac{1}{2}u_1, \text{ subject to } \begin{bmatrix} x & \mathbf{z}^H \\ \mathbf{z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}, \mathbf{z}_{\Omega} = \mathbf{y}_{\Omega}. \quad (132)$$

The following result shows that the frequencies can be exactly recovered by solving (132) if sufficiently many samples are observed and the same frequency separation condition as above is satisfied.

Theorem 6.5 ([127]). Suppose we observe $\mathbf{y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$ on the index set Ω , where $\Omega \subset \{1, \dots, N\}$ is of size M and is selected uniformly at random. Assume that $\{\phi_j\}_{j=1}^K$ are drawn i.i.d. from the uniform distribution on the complex unit circle.¹ If $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$, then there exists a numerical constant C such that

$$M \geq C \max \left\{ \log^2 \frac{N}{\delta}, K \log \frac{K}{\delta} \log \frac{N}{\delta} \right\} \quad (133)$$

is sufficient to guarantee that, with probability at least $1 - \delta$, \mathbf{y} is the unique optimizer for (132) and $\mathbf{y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$ is the unique atomic decomposition satisfying $\|\mathbf{y}\|_{\mathcal{A}} = \sum_{j=1}^K c_j$.

¹This condition has been relaxed in [20], where it was assumed that $\{\phi_j\}_{j=1}^K$ are independent with zero mean.

In the presence of noise, the SDP resulting from the unconstrained formulation in (116) is given by:

$$\min_{x, \mathbf{u}, \mathbf{z}} \frac{\lambda}{2} (x + u_1) + \frac{1}{2} \|\mathbf{z}_\Omega - \mathbf{y}_\Omega\|_2^2, \text{ subject to } \begin{bmatrix} x & \mathbf{z}^H \\ \mathbf{z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (134)$$

While it is clear that the regularization parameter λ is used to balance the signal sparsity and the data fidelity, it is less clear how to choose it. Under the assumption of i.i.d. Gaussian noise, this choice has been studied in [81, 128, 130]. For ULAs the paper [81] shows that if we let $\lambda \approx \sqrt{M \log M \sigma}$, where σ denotes the noise variance, then the signal estimate $\hat{\mathbf{z}}$ given by (134) has the following per-element expected reconstruction error:

$$\frac{1}{M} \mathbb{E} \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2 \leq \sqrt{\frac{\log M}{M}} \sigma \cdot \sum_{k=1}^K c_k. \quad (135)$$

This error bound implies that if $K = o\left(\sqrt{\frac{M}{\log M}}\right)$, then the estimate $\hat{\mathbf{z}}$ is statistically consistent. Moreover, the paper [128] shows that if we let $\lambda = C\sqrt{M \log M \sigma}$, where $C > 1$ is a constant (not explicitly given), and if the frequencies are sufficiently separated as in Theorem 6.4, then the following error bound can be obtained with high probability:

$$\frac{1}{M} \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2 = O\left(\frac{K \log M}{M} \sigma\right), \quad (136)$$

which is nearly minimax optimal. This implies that the estimate is consistent if $K = o\left(\frac{M}{\log M}\right)$. Furthermore, the frequencies and the amplitudes can be stably estimated as well [128]. Finally, note that the result in [81] has been generalized to the SLA case in [130]. It was shown that if we let $\lambda \approx \sqrt{M \log N \sigma}$ in such a case, it holds similarly to (135) that

$$\frac{1}{M} \mathbb{E} \|\hat{\mathbf{z}}_\Omega - \mathbf{z}_\Omega\|_2^2 \leq \sqrt{\frac{\log N}{M}} \sigma \cdot \sum_{k=1}^K c_k. \quad (137)$$

This means that the estimate $\hat{\mathbf{z}}_\Omega$ is consistent if $K = o\left(\sqrt{\frac{M}{\log N}}\right)$.

6.3.4 Hankel-based Nuclear Norm

Another choice of $\mathcal{M}(\mathbf{z})$ is the Hankel-based nuclear norm that was proposed in [129]. This metric is introduced based on the following observation. Given \mathbf{z} as in (114), let us form the Hankel matrix:

$$\mathbf{H}(\mathbf{z}) = \begin{bmatrix} z_1 & z_2 & \dots & z_n \\ z_2 & z_3 & \dots & z_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ z_m & z_{m+1} & \dots & z_N \end{bmatrix}, \quad (138)$$

where $m + n = N + 1$. It follows that

$$\mathbf{H}(\mathbf{z}) = \sum_{k=1}^K s_k \begin{bmatrix} 1 \\ e^{i2\pi f_k} \\ \vdots \\ e^{i2\pi(m-1)f_k} \end{bmatrix} [1 \quad e^{i2\pi f_k} \quad \dots \quad e^{i2\pi(n-1)f_k}]. \quad (139)$$

If $K < \min(m, n)$, then we have that $\mathbf{H}(\mathbf{z})$ is a low rank matrix with

$$\text{rank}(\mathbf{H}(\mathbf{z})) = K. \quad (140)$$

To reconstruct \mathbf{z} , therefore, we may consider the reconstruction of $\mathbf{H}(\mathbf{z})$ by choosing the sparse metric as $\text{rank}(\mathbf{H}(\mathbf{z}))$. If \mathbf{z} can be determined for the resulting rank minimization problem, then the frequencies may be recovered from \mathbf{z} .

Since the rank minimization cannot be easily solved, we seek a convex relaxation of $\text{rank}(\mathbf{H}(\mathbf{z}))$. The nuclear norm is a natural choice, which leads to:

$$\mathcal{M}(\mathbf{z}) = \|\mathbf{H}(\mathbf{z})\|_*. \quad (141)$$

The optimization problems resulting from (115)-(117) by using (141) are referred to as enhanced matrix completion (EMaC) in [129]. Note that the nuclear norm can be formulated as the following SDP [144]:

$$\|\mathbf{H}(\mathbf{z})\|_* = \min_{\mathbf{Q}_1, \mathbf{Q}_2} \frac{1}{2} [\text{Tr}(\mathbf{Q}_1) + \text{Tr}(\mathbf{Q}_2)], \text{ subject to } \begin{bmatrix} \mathbf{Q}_1 & \mathbf{H}(\mathbf{z})^H \\ \mathbf{H}(\mathbf{z}) & \mathbf{Q}_2 \end{bmatrix} \geq \mathbf{0}. \quad (142)$$

As a result, like the atomic norm method, the EMaC problems can be cast as SDP and solved using off-the-shelf solvers.

Theoretical guarantees for EMaC have been provided in the SLA case in [129] which, to some extent, are similar to those for the atomic norm method. In particular, it was shown that the signal \mathbf{z} can be exactly recovered in the absence of noise and stably recovered in the presence of bounded noise if the number of measurements M exceeds a constant times the number of sinusoids K up to a polylog factor, as given by (133), and if a certain coherence condition is satisfied. It is argued in [129] that the coherence condition required by EMaC can be weaker than the frequency separation condition required by ANM and thus higher resolution might be obtained by EMaC as compared to ANM. Connections between the two methods will be studied in the following subsection.

6.3.5 Connection between ANM and EMaC

To investigate the connection between ANM and EMaC, we define the following set of complex exponentials as a new atomic set:

$$\mathcal{A}' := \left\{ \mathbf{a}'(\phi) = [1, \phi, \dots, \phi^{N-1}]^T : \phi \in \mathbb{C} \right\}. \quad (143)$$

It is evident that, as compared to \mathcal{A}' , the complex exponentials are restricted to have constant modulus in \mathcal{A} with $|\phi| = 1$ and thus \mathcal{A} is a subset of \mathcal{A}' . For any $\mathbf{z} \in \mathbb{C}^N$, we can similarly define the atomic ℓ_0 norm with respect to \mathcal{A}' , denoted by $\|\mathbf{z}\|_{\mathcal{A}',0}$. We have the following result.

Theorem 6.6. *For any \mathbf{z} it holds that*

$$\|\mathbf{z}\|_{\mathcal{A}',0} \geq \text{rank}(\mathbf{H}(\mathbf{z})). \quad (144)$$

Moreover, if $\text{rank}(\mathbf{H}(\mathbf{z})) < \min(m, n)$, then

$$\|\mathbf{z}\|_{\mathcal{A}',0} = \text{rank}(\mathbf{H}(\mathbf{z})) \quad (145)$$

except for degenerate cases.

Proof. Suppose that $\|\mathbf{z}\|_{\mathcal{A}',0} = K'$. This means that there exists a K' -atomic decomposition for \mathbf{z} with respect to \mathcal{A}' :

$$\mathbf{z} = \sum_{k=1}^{K'} \mathbf{a}(\phi_k) s'_k, \quad \phi_k \in \mathbb{C}. \quad (146)$$

It follows that $\mathbf{H}(\mathbf{z})$ admits a decomposition similar to (139) and therefore that $\text{rank}(\mathbf{H}(\mathbf{z})) \leq K'$ and hence (144) holds.

The second part can be shown by applying the Kronecker's theorem for Hankel matrices (see, e.g., [145]). In particular, the Kronecker's theorem states that if $\text{rank}(\mathbf{H}(z)) = K' < \min(m, n)$, then z can be written as in (146) except for degenerate cases. According to the definition of $\|z\|_{\mathcal{A}',0}$, we have that

$$\|z\|_{\mathcal{A}',0} \leq K' = \text{rank}(\mathbf{H}(z)) \quad (147)$$

This together with (144) concludes the proof of (145). \blacksquare

By Theorem 6.6, we have linked $\text{rank}(\mathbf{H}(z))$, which motivated the use of its convex relaxation $\|\mathbf{H}(z)\|_*$, to an atomic ℓ_0 norm. In the regime of interest here $\mathbf{H}(z)$ is low-rank and hence $\text{rank}(\mathbf{H}(z))$ is almost identical to the atomic ℓ_0 norm induced by \mathcal{A}' . To compare $\|z\|_{\mathcal{A},0}$ and $\|z\|_{\mathcal{A}',0}$, we have the following result.

Theorem 6.7. *For any z it holds that*

$$\|z\|_{\mathcal{A}',0} \leq \|z\|_{\mathcal{A},0}. \quad (148)$$

Proof. The inequality is a direct consequence of the fact that $\mathcal{A} \subset \mathcal{A}'$. \blacksquare

By Theorem 6.7 the newly defined $\|z\|_{\mathcal{A}',0}$, which is closely related to $\text{rank}(\mathbf{H}(z))$, is actually a lower bound on $\|z\|_{\mathcal{A},0}$. It is worth noting that this lower bound is obtained by ignoring a known structure of the signal: from \mathcal{A} to \mathcal{A}' we have neglected the prior knowledge that each exponential component of z has constant modulus. As a consequence, using $\|z\|_{\mathcal{A}',0}$ as the sparse metric instead of $\|z\|_{\mathcal{A},0}$, we cannot guarantee in general, especially in the noisy case, that each component of the obtained signal z corresponds to one frequency. Note that this is also true for the convex relaxation metric $\|\mathbf{H}(z)\|_*$ as compared to $\|z\|_{\mathcal{A}}$. In contrast to this, the frequencies can be directly retrieved from the solution of the atomic norm method. From this point of view, the atomic norm method may be expected to outperform EMaC due to its better capability to capture the signal structure.

On the other hand, EMaC might have higher resolution than the atomic norm method. This is indeed true in the noiseless ULA case where the signal z is completely known. In this extreme case EMaC does not suffer from any resolution limit, while the atomic norm requires a frequency separation condition for its successful operation (at least theoretically).

6.3.6 Covariance Fitting Method: Gridless SPICE (GLS)

GLS was introduced in [130, 136] as a gridless version of the SPICE method presented in Subsection 4.5. Since SPICE is covariance-based and the data covariance matrix is a highly nonlinear function of the DOA parameters of interest, gridding is performed in SPICE to linearize the problem based on the zeroth order approximation. But this is not required in the case of ULAs or SLAs. The key idea of GLS is to re-parameterize the data covariance matrix using a PSD Toeplitz matrix $\mathbf{T}(\mathbf{u})$, which is linear in the new parameter vector \mathbf{u} , by making use of the Vandermonde decomposition of Toeplitz covariance matrices. To derive GLS, naturally, we make the same assumptions as for SPICE.

We first consider the ULA case. Assume that the noise is homoscedastic (note that, like SPICE, GLS can be extended to the case of heteroscedastic noise). It follows from the arguments in Subsection 6.2 that the data covariance matrix \mathbf{R} is a Toeplitz matrix. Therefore, \mathbf{R} can be linearly re-parameterized as:

$$\mathbf{R} = \mathbf{T}(\mathbf{u}), \quad \mathbf{T}(\mathbf{u}) \geq \mathbf{0}. \quad (149)$$

For a single snapshot, SPICE minimizes the following covariance fitting criterion:

$$\left\| \mathbf{R}^{-\frac{1}{2}} (\mathbf{y}\mathbf{y}^H - \mathbf{R}) \right\|_{\text{F}}^2 = \|\mathbf{y}\|_2^2 \cdot \mathbf{y}^H \mathbf{R}^{-1} \mathbf{y} + \text{Tr}(\mathbf{R}) - 2 \|\mathbf{y}\|_2^2. \quad (150)$$

Inserting (149) into (150), the resulting GLS optimization problem is given by:

$$\begin{aligned}
& \min_{\mathbf{u}} \|\mathbf{y}\|_2^2 \cdot \mathbf{y}^H \mathbf{T}^{-1}(\mathbf{u}) \mathbf{y} + \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\
& \Leftrightarrow \min_{x, \mathbf{u}} \|\mathbf{y}\|_2^2 x + Mu_1, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \text{ and } x \geq \mathbf{y}^H \mathbf{T}^{-1}(\mathbf{u}) \mathbf{y} \\
& \Leftrightarrow \min_{x, \mathbf{u}} \|\mathbf{y}\|_2^2 x + Mu_1, \text{ subject to } \begin{bmatrix} x & \mathbf{y}^H \\ \mathbf{y} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}.
\end{aligned} \tag{151}$$

Therefore, the covariance fitting problem has been cast as an SDP that can be solved in a polynomial time. Once the problem is solved, the data covariance estimate $\hat{\mathbf{R}} = \mathbf{T}(\hat{\mathbf{u}})$ is obtained, where $\hat{\mathbf{u}}$ denotes the solution of \mathbf{u} . Finally, the estimates of the parameters $(\hat{\mathbf{f}}, \hat{\mathbf{p}}, \hat{\sigma})$ can be obtained from the decomposition of $\hat{\mathbf{R}}$ by applying Corollary 6.1. Moreover, we note that the GLS optimization problem in (151) is very similar to the SDP for the atomic norm. Their connection will be discussed later in more detail.

In the SLA case, corresponding to (149), the covariance fitting criterion of SPICE is given by:

$$\left\| \mathbf{R}_{\Omega}^{-\frac{1}{2}} (\mathbf{y}_{\Omega} \mathbf{y}_{\Omega}^H - \mathbf{R}_{\Omega}) \right\|_{\text{F}}^2 = \|\mathbf{y}_{\Omega}\|_2^2 \cdot \mathbf{y}_{\Omega}^H \mathbf{R}_{\Omega}^{-1} \mathbf{y}_{\Omega} + \text{Tr}(\mathbf{R}_{\Omega}) - 2 \|\mathbf{y}_{\Omega}\|_2^2, \tag{152}$$

where \mathbf{R}_{Ω} denotes the covariance matrix of \mathbf{y}_{Ω} . To explicitly express \mathbf{R}_{Ω} , we let $\mathbf{\Gamma}_{\Omega} \in \{0, 1\}^{M \times N}$ be the row-selection matrix satisfying

$$\mathbf{y}_{\Omega} = \mathbf{\Gamma}_{\Omega} \mathbf{y}, \tag{153}$$

where $\mathbf{y} \in \mathbb{C}^N$ denotes the full data vector of the virtual N -element ULA. More concretely, $\mathbf{\Gamma}_{\Omega}$ is such that its j th row contains all 0s but a single 1 at the Ω_j th position. It follows that

$$\mathbf{R}_{\Omega} = \mathbb{E} \mathbf{y}_{\Omega} \mathbf{y}_{\Omega}^H = \mathbf{\Gamma}_{\Omega} \cdot \mathbb{E} \mathbf{y} \mathbf{y}^H \cdot \mathbf{\Gamma}_{\Omega}^T = \mathbf{\Gamma}_{\Omega} \mathbf{R} \mathbf{\Gamma}_{\Omega}^T. \tag{154}$$

This means that \mathbf{R}_{Ω} is a submatrix of the covariance matrix \mathbf{R} of the virtual full data \mathbf{y} . Therefore, using the parameterization of \mathbf{R} as in (149), \mathbf{R}_{Ω} can be linearly parameterized as:

$$\mathbf{R}_{\Omega} = \mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T, \quad \mathbf{T}(\mathbf{u}) \geq \mathbf{0}. \tag{155}$$

Inserting (155) into (152), we have that the GLS optimization problem resulting from (152) can be cast as the following SDP:

$$\min_{x, \mathbf{u}} \|\mathbf{y}_{\Omega}\|_2^2 x + Mu_1, \text{ subject to } \begin{bmatrix} x & \mathbf{y}_{\Omega}^H \\ \mathbf{y}_{\Omega} & \mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T \end{bmatrix} \geq \mathbf{0}. \tag{156}$$

Once the SDP is solved, the parameters of interest can be retrieved from the full data covariance estimate $\hat{\mathbf{R}} = \mathbf{T}(\hat{\mathbf{u}})$ as in the ULA case.

GLS is guaranteed to produce a sparse solution with at most $N - 1$ sources. This is a direct consequence of the frequency retrieval step, see Corollary 6.1. In general, GLS overestimates the true source number K in the presence of noise. This is reasonable because in GLS we do not assume any knowledge of the source number or of the noise variance(s).

An automatic source number estimation approach (a.k.a. model order selection) has been proposed in [130] based on the eigenvalues of the data covariance estimate $\hat{\mathbf{R}}$. The basic intuition behind it is that the larger eigenvalues correspond to the sources while the smaller ones are caused more likely by noise. The SORT algorithm [146] is adopted in [130] to identify these two groups of eigenvalues and thus provide an estimate of the source number. Furthermore, based on the source number, the frequency estimates can be refined by using a subspace method such as MUSIC. Readers are referred to [130] for details.

6.3.7 Connection between ANM and GLS

GLS is strongly connected to ANM. In this subsection, we show that GLS is equivalent to ANM as if there were no noise and with a slightly different frequency retrieval process [130]. We note that a similar connection in the discrete setting has been provided in [88, 89].

In the ULA case this connection can be shown based on the following equivalences/equalities:

$$\begin{aligned}
(151) &\Leftrightarrow \min_{\mathbf{u}} \|\mathbf{y}\|_2^2 \cdot \mathbf{y}^H \mathbf{T}^{-1}(\mathbf{u}) \mathbf{y} + M u_1, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\
&\Leftrightarrow \min_{\mathbf{u}} M \left\{ \left[\frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y}^H \right] \mathbf{T}^{-1}(\mathbf{u}) \left[\frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y} \right] + u_1 \right\}, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\
&\Leftrightarrow \min_{x, \mathbf{u}} M \{x + u_1\}, \text{ subject to } \begin{bmatrix} x & \frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y}^H \\ \frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0} \\
&\Leftrightarrow 2M \left\| \frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y} \right\|_{\mathcal{A}} \\
&\Leftrightarrow 2\sqrt{M} \|\mathbf{y}\|_2 \cdot \|\mathbf{y}\|_{\mathcal{A}}.
\end{aligned} \tag{157}$$

This means that GLS is equivalent to computing the atomic norm of the noisy data \mathbf{y} (up to a scaling factor).

In the SLA case, we need the following equality that holds for $\mathbf{R} > \mathbf{0}$ [130]:

$$\mathbf{y}_{\Omega}^H [\mathbf{\Gamma}_{\Omega} \mathbf{R} \mathbf{\Gamma}_{\Omega}^T]^{-1} \mathbf{y}_{\Omega} = \min_{\mathbf{z}} \mathbf{z}^H \mathbf{R}^{-1} \mathbf{z}, \text{ subject to } \mathbf{z}_{\Omega} = \mathbf{y}_{\Omega}. \tag{158}$$

As a result, we have that

$$\begin{aligned}
(156) &\Leftrightarrow \min_{\mathbf{u}} \|\mathbf{y}_{\Omega}\|_2^2 \cdot \mathbf{y}_{\Omega}^H [\mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T]^{-1} \mathbf{y}_{\Omega} + M u_1, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\
&\Leftrightarrow \min_{\mathbf{u}, \mathbf{z}} \|\mathbf{y}_{\Omega}\|_2^2 \cdot \mathbf{z}^H \mathbf{T}^{-1}(\mathbf{u}) \mathbf{z} + M u_1, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0}, \mathbf{z}_{\Omega} = \mathbf{y}_{\Omega} \\
&\Leftrightarrow \min_{\mathbf{u}, \mathbf{z}} M \left\{ \left[\frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} \mathbf{z}^H \right] \mathbf{T}^{-1}(\mathbf{u}) \left[\frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} \mathbf{z} \right] + u_1 \right\}, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0}, \mathbf{z}_{\Omega} = \mathbf{y}_{\Omega} \\
&\Leftrightarrow \min_{x, \mathbf{u}, \mathbf{z}} M \{x + u_1\}, \text{ subject to } \begin{bmatrix} x & \frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} \mathbf{z}^H \\ \frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} \mathbf{z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}, \mathbf{z}_{\Omega} = \mathbf{y}_{\Omega} \\
&\Leftrightarrow \min_{\mathbf{z}} 2M \left\| \frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} \mathbf{z} \right\|_{\mathcal{A}}, \text{ subject to } \mathbf{z}_{\Omega} = \mathbf{y}_{\Omega} \\
&\Leftrightarrow \min_{\mathbf{z}} 2\sqrt{M} \|\mathbf{y}_{\Omega}\|_2 \cdot \|\mathbf{z}\|_{\mathcal{A}}, \text{ subject to } \mathbf{z}_{\Omega} = \mathbf{y}_{\Omega} \\
&\Leftrightarrow \min_{\mathbf{z}} \|\mathbf{z}\|_{\mathcal{A}}, \text{ subject to } \mathbf{z}_{\Omega} = \mathbf{y}_{\Omega}.
\end{aligned} \tag{159}$$

This means, like in the ULA case, that GLS is equivalent to ANM subject to the data consistency as if there were no noise.

Finally, note that GLS is practically attractive since it does not require knowledge of the noise level. Regarding this fact, note that GLS is different from ANM in the frequency retrieval process: whereas Theorem 6.1 is used in ANM, Corollary 6.1 is employed in GLS since the noise variance is also encoded in the Toeplitz covariance matrix $\mathbf{T}(\mathbf{u})$, besides the frequencies.

6.4 The Multiple Snapshot Case: Covariance Fitting Methods

In this and the following subsections we will study gridless DOA estimation methods for multiple snapshots. The methods that we will introduce are mainly based on or inspired by those in the single snapshot case in the

preceding subsection. The key techniques of these methods exploit the temporal redundancy of the multiple snapshots for possibly improved performance. We have decided to introduce the covariance fitting methods first since they appeared earlier than their deterministic peers. In this kind of methods, differently from the deterministic ones, certain statistical assumptions on the sources (like for SPICE) are required to explicitly express the data covariance matrix. We will discuss three covariance-based gridless sparse methods: GLS in [136], the SMV-based atomic norm method in [147] and the low rank matrix denoising approach in [148]. While GLS is applicable to an arbitrary number of snapshots, the latter two can only be used if there are sufficiently many snapshots.

6.4.1 Gridless SPICE (GLS)

In the presence of multiple snapshots, GLS is derived in a similar way as in the single snapshot case by utilizing the convex reparameterization of the data covariance matrix \mathbf{R} in (149). For convenience, some derivations of SPICE provided in Subsection 4.5 will be repeated here. We first consider the ULA case. Let $\tilde{\mathbf{R}} = \frac{1}{L} \mathbf{Y} \mathbf{Y}^H$ be the sample covariance. In the case of $L \geq M$ when the sample covariance $\tilde{\mathbf{R}}$ is nonsingular, the following SPICE covariance fitting criterion is minimized:

$$\left\| \mathbf{R}^{-\frac{1}{2}} (\tilde{\mathbf{R}} - \mathbf{R}) \tilde{\mathbf{R}}^{-\frac{1}{2}} \right\|_{\text{F}}^2 = \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}) + \text{Tr}(\tilde{\mathbf{R}}^{-1} \mathbf{R}) - 2M. \quad (160)$$

Inserting (149) into (160), we have the following equivalences:

$$\begin{aligned} & \min_{\mathbf{u}} \text{Tr} \left(\tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{T}^{-1}(\mathbf{u}) \tilde{\mathbf{R}}^{\frac{1}{2}} \right) + \text{Tr} \left(\tilde{\mathbf{R}}^{-1} \mathbf{T}(\mathbf{u}) \right), \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\ \Leftrightarrow & \min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr} \left(\tilde{\mathbf{R}}^{-1} \mathbf{T}(\mathbf{u}) \right), \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \text{ and } \mathbf{X} \geq \tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{T}^{-1}(\mathbf{u}) \tilde{\mathbf{R}}^{\frac{1}{2}} \\ \Leftrightarrow & \min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr} \left(\tilde{\mathbf{R}}^{-1} \mathbf{T}(\mathbf{u}) \right), \text{ subject to } \begin{bmatrix} \mathbf{X} & \tilde{\mathbf{R}}^{\frac{1}{2}} \\ \tilde{\mathbf{R}}^{\frac{1}{2}} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \end{aligned} \quad (161)$$

Therefore, the covariance fitting problem is cast as SDP. Once the problem is solved, the estimates of the parameters $(\hat{\mathbf{f}}, \hat{\mathbf{p}}, \hat{\sigma})$ can be obtained from the data covariance estimate $\hat{\mathbf{R}} = \mathbf{T}(\hat{\mathbf{u}})$ by applying Corollary 6.1.

In the case of $L < M$ when $\tilde{\mathbf{R}}$ is singular, we instead minimize the criterion

$$\left\| \mathbf{R}^{-\frac{1}{2}} (\tilde{\mathbf{R}} - \mathbf{R}) \right\|_{\text{F}}^2 = \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}^2) + \text{Tr}(\mathbf{R}) - 2\text{Tr}(\tilde{\mathbf{R}}). \quad (162)$$

Similarly, inserting (149) into (162), we obtain the following SDP:

$$\min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \tilde{\mathbf{R}} \\ \tilde{\mathbf{R}} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (163)$$

The parameter estimates $(\hat{\mathbf{f}}, \hat{\mathbf{p}}, \hat{\sigma})$ can be obtained in the same manner as above.

The dimensionality of the SDP in (163) can be reduced [82]. To do so, let $\tilde{\mathbf{Y}} = \frac{1}{L} \mathbf{Y} (\mathbf{Y}^H \mathbf{Y})^{\frac{1}{2}} \in \mathbb{C}^{M \times L}$ and observe that

$$\tilde{\mathbf{R}}^2 = \frac{1}{L^2} \mathbf{Y} \mathbf{Y}^H \mathbf{Y} \mathbf{Y}^H = \tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^H. \quad (164)$$

Inserting (164) into (162), we obtain another SDP formulation of the covariance fitting problem:

$$\min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \tilde{\mathbf{Y}}^H \\ \tilde{\mathbf{Y}} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (165)$$

Compared to (163), the dimensionality of the semidefinite matrix in (165) is reduced from $2M \times 2M$ to $(M + L) \times (M + L)$ (note that in this case $L < M$). This reduction can be significant in the case of a small number of snapshots.

Similar results can be obtained in the SLA case. In particular, let \mathbf{R}_Ω and $\tilde{\mathbf{R}}_\Omega = \frac{1}{L} \mathbf{Y}_\Omega \mathbf{Y}_\Omega^H$ denote the data covariance and the sample covariance, respectively. Using the linear reparameterization of \mathbf{R}_Ω in (155), similar SDPs to those above can be formulated. In the case of $L \geq M$ when $\tilde{\mathbf{R}}_\Omega$ is nonsingular, we have the following SDP:

$$\min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{\Gamma}_\Omega^T \tilde{\mathbf{R}}_\Omega^{-1} \mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \tilde{\mathbf{R}}_\Omega^{\frac{1}{2}} \\ \tilde{\mathbf{R}}_\Omega^{\frac{1}{2}} & \mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_\Omega^T \end{bmatrix} \geq \mathbf{0}. \quad (166)$$

When $L < M$, the SDP is

$$\min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{\Gamma}_\Omega^T \mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \tilde{\mathbf{R}}_\Omega \\ \tilde{\mathbf{R}}_\Omega & \mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_\Omega^T \end{bmatrix} \geq \mathbf{0}, \quad (167)$$

where $\tilde{\mathbf{R}}_\Omega$ can also be replaced by $\frac{1}{L} \mathbf{Y}_\Omega (\mathbf{Y}_\Omega^H \mathbf{Y}_\Omega)^{\frac{1}{2}} \in \mathbb{C}^{M \times L}$ for dimensionality reduction. Once the SDP is solved, the parameters of interest can be retrieved as in the ULA case.

As in the single snapshot case, GLS is guaranteed to produce a sparse solution with at most $N - 1$ sources. Besides this, GLS has other attractive properties as detailed below.

Let us assume that the antenna array is a redundancy array or, equivalently, that the full matrix $\mathbf{T}(\mathbf{u})$ can be recovered from its submatrix $\mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_\Omega^T$. Note that all ULAs and many SLAs are redundancy arrays. Then the GLS estimator is statistically consistent in the snapshot number L if the true source number $K \leq N - 1$. To see this, let us consider a ULA first. As $L \rightarrow \infty$, $\tilde{\mathbf{R}}$ approaches the true data covariance matrix that is denoted by \mathbf{R}^o and has the same Toeplitz structure as \mathbf{R} . Hence, it follows from (160) that $\hat{\mathbf{R}} = \mathbf{R}^o$ and the true parameters can be retrieved from $\hat{\mathbf{R}}$. In the SLA case, similarly, the covariance matrix estimate $\hat{\mathbf{R}}_\Omega$ converges to the true one as $L \rightarrow \infty$. Then, the assumption of redundancy array can be used to show that all the information in the Toeplitz matrix $\mathbf{T}(\mathbf{u})$ for frequency retrieval can be recovered from $\hat{\mathbf{R}}_\Omega$ and hence that the true parameters can be obtained.

Furthermore, under the stronger assumption of Gaussian sources and noise, GLS is an asymptotic ML estimator when $K \leq N - 1$ and a redundancy array is employed. This is true because the global solution of the SPICE covariance fitting problem is a large-snapshot realization of the ML estimator ([85, 149]) and because GLS globally solves the problem. As a direct consequence of this, GLS has improved resolution as L increases and the resolution limit vanishes as L grows to infinity. Another consequence is that GLS can estimate more sources than the number of antennas. In fact, a redundancy array with aperture N can usually be formed by using $M \approx \sqrt{3(N - 1)}$ antennas [150]. This means that up to about $\frac{1}{3}M^2$ sources can be estimated using GLS with only M antennas.

It is worth noting that generally the above properties of GLS are not shared by its discrete version, viz. SPICE, due to an ambiguity problem, even when the on-grid assumption of SPICE holds. To see this, let us consider the ULA case as an example and suppose that SPICE can accurately estimate the data covariance matrix $\mathbf{R} = \mathbf{T}(\mathbf{u})$, as GLS does. Note that when \mathbf{R} has full rank, which is typically the case in practice, the solution of GLS is provided by the unique decomposition of \mathbf{R} from Corollary 6.1. But this uniqueness cannot be guaranteed in SPICE according to Remark 6.1 (note that the condition that $r < N$ of Corollary 6.1 might not be satisfied in SPICE).

6.4.2 SMV-based Atomic Norm Minimization (ANM-SMV)

Within the SMV super-resolution framework introduced in [97], an ANM approach was proposed in [147], designated here as ANM-SMV. While the paper [147] focused on co-prime arrays ([151]), which form a special class of SLAs, ANM-SMV can actually be applied to a broader class of SLAs such as redundancy arrays or even general SLAs. To simplify our discussions, we consider without loss of generality a redundancy array, denoted by Ω .

Under the assumption of uncorrelated sources, as for GLS, the data covariance matrix \mathbf{R}_Ω can be expressed as:

$$\mathbf{R}_\Omega = \mathbf{A}_\Omega(\mathbf{f}) \text{diag}(\mathbf{p}) \mathbf{A}_\Omega^H(\mathbf{f}) + \sigma \mathbf{I}, \quad (168)$$

where $p_k > 0$, $k = 1, \dots, K$ denote the source powers. According to the discussions in Subsection 6.3.6, \mathbf{R}_Ω is a submatrix of a Toeplitz covariance matrix:

$$\mathbf{R}_\Omega = \mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_\Omega^T + \sigma \mathbf{I}, \quad (169)$$

where $\mathbf{u} \in \mathbb{C}^N$ and $\mathbf{T}(\mathbf{u}) = \mathbf{A}(\mathbf{f}) \text{diag}(\mathbf{p}) \mathbf{A}^H(\mathbf{f})$. Since the frequencies have been encoded in $\mathbf{T}(\mathbf{u})$, ANM-SMV, like GLS, also carries out covariance fitting to estimate \mathbf{u} and thus the frequencies. But the technique used by ANM-SMV is different. To describe this technique, note that \mathbf{u} can be expressed as

$$\mathbf{u} = \mathbf{A}^*(\mathbf{f}) \mathbf{p}. \quad (170)$$

Let us define $\mathbf{v} \in \mathbb{C}^{2N-1}$ such that

$$v_j = \begin{cases} u_{N-j+1}, & j = 1, \dots, N, \\ u_{j-N+1}^*, & j = N+1, \dots, 2N-1. \end{cases} \quad (171)$$

Given \mathbf{u} in (170), note that \mathbf{v} can be viewed as a snapshot of a virtual $(2N-1)$ -element ULA on which K sources impinge. Based on this observation, ANM-SMV attempts to solve the following ANM problem:

$$\min_{\mathbf{v}} \|\mathbf{v}\|_{\mathcal{A}}, \text{ subject to } \|\mathbf{v} - \tilde{\mathbf{v}}\|_2 \leq \eta, \quad (172)$$

where $\tilde{\mathbf{v}}$ denotes an estimate of \mathbf{v} , which will be given later, and η is an upper bound on the error of $\tilde{\mathbf{v}}$. By casting (172) as SDP, this problem can be solved and the frequencies can then be estimated as those composing the solution \mathbf{v} .

The remaining task is to compose the estimate $\tilde{\mathbf{v}}$ and analyze its error bound η . To do so, the noise variance σ is assumed to be known. Using (169), an estimate of $\mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_\Omega^T$ is formed as $\tilde{\mathbf{R}}_\Omega - \sigma \mathbf{I}$. After that, an estimate of \mathbf{u} is obtained by averaging the corresponding entries of the estimate $\tilde{\mathbf{R}}_\Omega - \sigma \mathbf{I}$. This can be done since Ω is assumed to be a redundancy array. Finally, $\tilde{\mathbf{v}}$ is obtained by using (171). Under the assumption of i.i.d. Gaussian sources and noise and assuming sufficiently many snapshots, an error bound $\eta \propto \sigma$ is provided in [147] in the case of co-prime arrays. This bound might be extended to other cases but that is beyond the scope of this article. Based on the above observations and the result in [152], it can be shown that ANM-SMV can stably estimate the frequencies, provided that they are sufficiently separated, with a probability that increases with the snapshot number L .

6.4.3 Nuclear Norm Minimization followed by MUSIC (NNM-MUSIC)

Using a low rank matrix recovery technique and a subspace method, another covariance-based method was proposed in [148] that is called NNM-MUSIC. Based on (169), the paper [148] proposed a two-step approach: 1) First estimate the full data covariance matrix $\mathbf{T}(\mathbf{u})$ by exploiting its low rank structure, and 2) Then estimate the frequencies from $\mathbf{T}(\mathbf{u})$ using MUSIC.

In the first step, the following NNM problem is solved to estimate $\mathbf{T}(\mathbf{u})$:

$$\min_{\mathbf{R}} \|\mathbf{R}\|_*, \text{ subject to } \|\mathbf{R} - \mathbf{T}(\tilde{\mathbf{u}})\|_F \leq \eta. \quad (173)$$

In (173), $\mathbf{T}(\tilde{\mathbf{u}})$ denotes an estimate of the data covariance matrix, which is obtained by averaging the corresponding entries of the sample covariance matrix $\tilde{\mathbf{R}}_\Omega$, and η measures the distance between the true low rank matrix $\mathbf{T}(\mathbf{u})$ and the above estimate in the Frobenius norm. Once \mathbf{R} is solved for, MUSIC is adopted to estimate the frequencies from the subspace of \mathbf{R} .

By setting $\eta = \sqrt{N}\sigma$ and assuming $L \rightarrow \infty$, it was shown in [148] that solving (173) can exactly recover $\mathbf{T}(\mathbf{u})$. However, we note that it is not an easy task to choose η in practice. Moreover, although the Toeplitz structure embedded in the data covariance matrix \mathbf{R}_Ω is exploited to form the estimate $\tilde{\mathbf{u}}$, this structure is not utilized in the matrix denoising step. It was argued in [148] that the PSDness of \mathbf{R} can be preserved by solving (173) if $\mathbf{T}(\tilde{\mathbf{u}})$ is PSD, which holds true if sufficiently many snapshots are available.

6.4.4 Comparison of GLS, ANM-SMV and NNM-MUSIC

In this subsection we compare the three covariance-based methods, namely GLS, ANM-SMV and NNM-MUSIC. While these methods are derived under similar assumptions on sources and noise, we argue that GLS can outperform ANM-SMV and NNM-MUSIC in several aspects. First, GLS is hyperparameter-free and can consistently estimate the noise variance σ , whereas ANM-SMV and NNM-MUSIC usually require knowledge of this variance since the error bounds η in (172) and (173) are functions of σ . In fact, even when σ is known the choice of η is still not easy. Second, ANM-SMV and NNM-MUSIC are usable only with sufficiently many snapshots (which are needed to obtain a reasonable estimate of \mathbf{u} as well as a reasonable error bound η), while GLS can be used even with a single snapshot. Third, GLS and NNM-MUSIC are statistically consistent but ANM-SMV is not. Note that ANM-SMV still suffers from a resolution limit even if $\tilde{\mathbf{v}}$ in (172) is exactly estimated given an infinitely number of snapshots. Fourth, GLS is a large-snapshot realization of the ML estimation while ANM-SMV and NNM-MUSIC are not.

Last but not least, ANM-SMV and NNM-MUSIC cannot exactly recover the frequencies in the absence of noise since the estimate $\tilde{\mathbf{v}}$ or $\tilde{\mathbf{u}}$ will suffer from some approximation error with finite snapshots. In contrast to this, GLS can exactly recover the frequencies under a mild frequency separation condition due to its connection to the atomic norm that will be discussed in Subsection 6.5.3 (considering the single snapshot case as an example).

6.5 The Multiple Snapshot Case: Deterministic Methods

In this subsection we present several deterministic gridless sparse methods for the case of multiple snapshots. The main idea is to utilize the temporal redundancy of the snapshots. Different from the covariance-based methods in the preceding subsection, these deterministic optimization methods are derived without statistical assumptions on the sources (though weak technical assumptions might be needed to provide theoretical guarantees). As in the single snapshot case, we first provide a general framework for such methods. We then discuss the potential advantages of multiple snapshots for DOA/frequency estimation based on an MMV atomic ℓ_0 norm formulation. After that, the MMV atomic norm method will be presented. Finally, a possible extension of EMaC to the multiple snapshot case is discussed.

6.5.1 A General Framework

The data model in (99) can be written as:

$$\mathbf{Y}_\Omega = \mathbf{Z}_\Omega + \mathbf{E}_\Omega, \quad \mathbf{Z} := \mathbf{A}(\mathbf{f}) \mathbf{S}, \quad (174)$$

where \mathbf{Z} denotes the noiseless multiple snapshot signal that contains the frequencies of interest. In the presence of bounded noise with $\|\mathbf{E}_\Omega\|_F \leq \eta$, we solve the constrained optimization problem:

$$\min_{\mathbf{Z}} \mathcal{M}(\mathbf{Z}), \text{ subject to } \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F \leq \eta. \quad (175)$$

We can instead solve the regularized optimization problem given by

$$\min_{\mathbf{Z}} \lambda \mathcal{M}(\mathbf{Z}) + \frac{1}{2} \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F^2, \quad (176)$$

where $\lambda > 0$ is a regularization parameter. In the extreme noiseless case, both (175) and (176) degenerate to the following problem:

$$\min_{\mathbf{Z}} \mathcal{M}(\mathbf{Z}), \text{ subject to } \mathbf{Z}_\Omega = \mathbf{Y}_\Omega. \quad (177)$$

In (175)-(177), $\mathcal{M}(\mathbf{Z})$ denotes a sparse metric. By minimizing $\mathcal{M}(\mathbf{Z})$ we attempt to reduce the number of frequencies composing \mathbf{Z} .

6.5.2 Atomic ℓ_0 Norm

In this subsection we study the advantage of using multiple snapshots for frequency estimation and extend the result in Subsection 4.3.1 from the discrete to the continuous setting. To do so, we extend the atomic ℓ_0 norm from the single to the multiple snapshot case. Note that the noiseless multiple snapshot signal \mathbf{Z} in (174) can be written as:

$$\mathbf{Z} = \sum_{k=1}^K \mathbf{a}(f_k) \mathbf{s}_k = \sum_{k=1}^K c_k \mathbf{a}(f_k) \boldsymbol{\phi}_k, \quad (178)$$

where $\mathbf{s}_k := [s_{k1}, \dots, s_{kL}] \in \mathbb{C}^{1 \times L}$, $c_k = \|\mathbf{s}_k\|_2 > 0$, and $\boldsymbol{\phi}_k = c_k^{-1} \mathbf{s}_k \in \mathbb{C}^{1 \times L}$ with $\|\boldsymbol{\phi}_k\|_2 = 1$. We therefore define the set of atoms in this case as:

$$\mathcal{A} := \{\mathbf{a}(f_k, \boldsymbol{\phi}_k) := \mathbf{a}(f_k) \boldsymbol{\phi}_k : f_k \in \mathbb{T}, \boldsymbol{\phi}_k \in \mathbb{C}^{1 \times L}, \|\boldsymbol{\phi}_k\|_2 = 1\} \quad (179)$$

Note that \mathbf{Z} is a linear combination of K atoms in \mathcal{A} . The atomic ℓ_0 norm of \mathbf{Z} induced by the new atomic set \mathcal{A} is given by:

$$\begin{aligned} \|\mathbf{Z}\|_{\mathcal{A},0} &:= \inf_{c_k, f_k, \boldsymbol{\phi}_k} \left\{ \mathcal{K} : \mathbf{Z} = \sum_{k=1}^{\mathcal{K}} \mathbf{a}(f_k, \boldsymbol{\phi}_k) c_k, f_k \in \mathbb{T}, \|\boldsymbol{\phi}_k\|_2 = 1, c_k > 0 \right\} \\ &= \inf_{f_k, \mathbf{s}_k} \left\{ \mathcal{K} : \mathbf{Z} = \sum_{k=1}^{\mathcal{K}} \mathbf{a}(f_k) \mathbf{s}_k, f_k \in \mathbb{T} \right\}. \end{aligned} \quad (180)$$

Using the atomic ℓ_0 norm, in the noiseless case, the problem resulting from (177) is given by:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{\mathcal{A},0}, \text{ subject to } \mathbf{Z}_\Omega = \mathbf{Y}_\Omega. \quad (181)$$

To show the advantage of multiple snapshots, we define the continuous dictionary

$$\mathcal{A}_\Omega^1 := \{\mathbf{a}_\Omega(f) : f \in \mathbb{T}\} \quad (182)$$

and let $\text{spark}(\mathcal{A}_\Omega^1)$ be its spark. We have the following theoretical guarantee for (181) that generalizes the result in [18, Theorem 2.4].

Theorem 6.8 ([20]). $\mathbf{Y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$ is the unique solution to (181) if

$$K < \frac{\text{spark}(\mathcal{A}_\Omega^1) - 1 + \text{rank}(\mathbf{Y}_\Omega)}{2}. \quad (183)$$

Moreover, the atomic decomposition above is the only one that satisfies $\|\mathbf{Y}\|_{\mathcal{A},0} = K$.

Note that the condition in (183) coincides with that in (10) required to guarantee parameter identifiability for DOA estimation. In fact it can be shown that, in the noiseless case, the frequencies/DOAs can be uniquely determined by the atomic ℓ_0 norm optimization if and only if they can be uniquely identified (see, e.g., [153, Proposition 2]). Furthermore, the above result also holds true for general array geometries and even for general parameter estimation problems, provided that the atomic ℓ_0 norm is appropriately defined.

By Theorem 6.8 the frequencies can be exactly determined by (181) if the number of sources K is sufficiently small with respect to the array geometry Ω and the observed data \mathbf{Y}_Ω . Note that the number of recoverable frequencies can be increased, as compared to the SMV case, if $\text{rank}(\mathbf{Y}_\Omega) > 1$, which always happens but in the trivial case when the multiple snapshots in \mathbf{Y}_Ω are identical up to scaling factors.

As in the single snapshot case, computing $\|\mathbf{Z}\|_{\mathcal{A},0}$ can be cast as a rank minimization problem. To see this, let \mathbf{T} be a Toeplitz matrix and impose the condition that

$$\begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T} \end{bmatrix} \geq \mathbf{0}, \quad (184)$$

where \mathbf{X} is a free matrix variable. By invoking the Vandermonde decomposition, we see that \mathbf{T} admits a rank (\mathbf{T}) -atomic Vandermonde decomposition. Moreover, \mathbf{Z} lies in the range space of \mathbf{T} and thus it can be expressed by rank (\mathbf{T}) atoms. Formally, we have the following result.

Theorem 6.9 ([20, 154]). $\|\mathbf{Z}\|_{\mathcal{A},0}$ defined in (180) equals the optimal value of the following rank minimization problem:

$$\min_{\mathbf{X}, \mathbf{u}} \text{rank}(\mathbf{T}(\mathbf{u})), \text{ subject to (184)}. \quad (185)$$

While the rank minimization problem cannot be easily solved, we next discuss its convex relaxation, namely the atomic norm method.

6.5.3 Atomic Norm

To provide a practical approach, we study the atomic norm induced by the atomic set \mathcal{A} defined in (179). As in the single snapshot case, we have that

$$\begin{aligned} \|\mathbf{Z}\|_{\mathcal{A}} &:= \inf \{t > 0 : \mathbf{Z} \in t\text{conv}(\mathcal{A})\} \\ &= \inf_{c_k, f_k, \phi_k} \left\{ \sum_k c_k : \mathbf{Z} = \sum_k \mathbf{a}(f_k, \phi_k) c_k, f_k \in \mathbb{T}, \|\phi_k\|_2 = 1, c_k > 0 \right\} \\ &= \inf_{f_k, \mathbf{s}_k} \left\{ \sum_k \|\mathbf{s}_k\|_2 : \mathbf{Z} = \sum_k \mathbf{a}(f_k) \mathbf{s}_k, f_k \in \mathbb{T} \right\}. \end{aligned} \quad (186)$$

Note that $\|\mathbf{Z}\|_{\mathcal{A}}$ is in fact a continuous counterpart of the $\ell_{2,1}$ norm. Moreover, it is shown in the following result that $\|\mathbf{Z}\|_{\mathcal{A}}$ can also be cast as SDP.

Theorem 6.10 ([20, 83]). $\|\mathbf{Z}\|_{\mathcal{A}}$ defined in (186) equals the optimal value of the following SDP:

$$\min_{\mathbf{X}, \mathbf{u}} \frac{1}{2\sqrt{N}} [\text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u}))], \text{ subject to (184)}. \quad (187)$$

The proof of Theorem 6.10 is omitted since it is very similar to that in the case of a single snapshot. Similarly, the frequencies and the powers are encoded in the Toeplitz matrix $\mathbf{T}(\mathbf{u})$ and therefore, once $\mathbf{T}(\mathbf{u})$ is obtained, they can be retrieved from its Vandermonde decomposition.

By applying Theorem 6.10, in the noiseless case, the ANM problem resulting from (177) can be cast as the following SDP:

$$\min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to (184) and } \mathbf{Z}_{\Omega} = \mathbf{Y}_{\Omega}. \quad (188)$$

For (188), we have theoretical guarantees similar to those in the single snapshot case; in other words, the frequencies can be exactly recovered by solving (188) under appropriate conditions. Formally, we have the following results that generalize those in the single snapshot case.

Theorem 6.11 ([20]). $\mathbf{Y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$ is the unique atomic decomposition satisfying $\|\mathbf{Y}\|_{\mathcal{A}} = \sum_{j=1}^K c_j$ if $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$ and $N \geq 257$.

Theorem 6.12 ([20]). Suppose we observe the rows of $\mathbf{Y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$ that are indexed by Ω , where $\Omega \subset [1, \dots, N]$ is of size M and is selected uniformly at random. Assume that $\{\phi_j\}$ are independent random variables on the unit hyper-sphere with $\mathbb{E}\phi_j = \mathbf{0}$. If $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$, then there exists a numerical constant C such that

$$M \geq C \max \left\{ \log^2 \frac{\sqrt{LN}}{\delta}, K \log \frac{K}{\delta} \log \frac{\sqrt{LN}}{\delta} \right\} \quad (189)$$

is sufficient to guarantee that, with probability at least $1 - \delta$, \mathbf{Y} is the unique solution to (188) and $\mathbf{Y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$ is the unique atomic decomposition satisfying $\|\mathbf{Y}\|_{\mathcal{A}} = \sum_{j=1}^K c_j$.

Note that we have not made any assumptions on the sources in Theorem 6.11 and therefore it applies to all kinds of source signals, including coherent sources. As a result, one cannot expect that the theoretical guarantee improves over the single snapshot case.

Note that the dependence of M on L in the bound (189) is for controlling the probability of successful recovery. To make it clear, consider the case when we seek to recover the columns of \mathbf{Y} independently using the single snapshot ANM method. When M satisfies (189) for $L = 1$, each column of \mathbf{Y} can be recovered with probability $1 - \delta$. It follows that \mathbf{Y} can be exactly recovered with probability at least $1 - L\delta$. In contrast to this, if we recover \mathbf{Y} using the multiple snapshot ANM method, then with the same number of measurements the success probability is improved to $1 - \sqrt{L}\delta$ (to see this, replace δ in (189) by $\sqrt{L}\delta$).

Note also that the assumption on $\{\phi_j\}$ in Theorem 6.12 is weak in the sense that the sources can be coherent. To see this, suppose that the rows of \mathbf{S} are i.i.d. Gaussian with mean zero and a covariance matrix whose rank is equal to one. Then the sources are identical up to random global phases and hence they are independent but coherent. This explains why the theoretical guarantee given in Theorem 6.12 does not improve over the similar result in the single snapshot case. In other words, the results of Theorems 6.11 and 6.12 are *worst case* analysis. Although these results do not shed light on the advantage of multiple snapshots, numerical simulations show that the atomic norm approach significantly improves the recovery performance, compared to the case of $L = 1$, when the source signals are at general positions (see, e.g., [20, 83, 154]).

In the presence of noise, the ANM problem resulting from (176) can be formulated as the following SDP:

$$\min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \frac{\lambda}{2\sqrt{N}} [\text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u}))] + \frac{1}{2} \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_{\text{F}}^2, \text{ subject to (184)}. \quad (190)$$

It was shown in [83] that in the ULA case the choice of $\lambda \approx \sqrt{M(L + \log M + \sqrt{2L \log M})} \sigma$ results in a stable recovery of the signal \mathbf{Z} , which generalizes the result in the single snapshot case. In the SLA case, a similar parameter tuning method can be derived by combining the techniques in [83] and [130].

6.5.4 Hankel-based Nuclear Norm

In this subsection we extend the EMaC method from the single to the multiple snapshot case. For each noiseless snapshot $\mathbf{z}(t)$, we can form an $m \times n$ Hankel matrix $\mathbf{H}(\mathbf{z}(t))$ as in (138), where $m + n = N + 1$, and then combine them in the following $m \times nL$ matrix:

$$\mathbf{H}(\mathbf{Z}) := [\mathbf{H}(\mathbf{z}(1)), \dots, \mathbf{H}(\mathbf{z}(L))]. \quad (191)$$

Using the decomposition in (139), we have that

$$\begin{aligned} \mathbf{H}(\mathbf{Z}) = \sum_{k=1}^K \begin{bmatrix} 1 \\ e^{i2\pi f_k} \\ \vdots \\ e^{i2\pi(m-1)f_k} \end{bmatrix} \times \\ \left[s_k(1), s_k(1)e^{i2\pi f_k}, \dots, s_k(1)e^{i2\pi(n-1)f_k}, \dots, s_k(L), s_k(L)e^{i2\pi f_k}, \dots, s_k(L)e^{i2\pi(n-1)f_k} \right]. \end{aligned} \quad (192)$$

As a result,

$$\text{rank}(\mathbf{H}(\mathbf{Z})) \leq K. \quad (193)$$

It follows that if $K < \min(m, nL)$, then $\mathbf{H}(\mathbf{Z})$ is a low rank matrix. Therefore, we may recover $\mathbf{H}(\mathbf{Z})$ by minimizing its nuclear norm, i.e., by letting (see (142))

$$\begin{aligned} \mathcal{M}(\mathbf{Z}) &= \|\mathbf{H}(\mathbf{Z})\|_* \\ &= \min_{\mathbf{Q}_1, \mathbf{Q}_2} \frac{1}{2} [\text{Tr}(\mathbf{Q}_1) + \text{Tr}(\mathbf{Q}_2)], \text{ subject to } \begin{bmatrix} \mathbf{Q}_1 & \mathbf{H}(\mathbf{Z})^H \\ \mathbf{H}(\mathbf{Z}) & \mathbf{Q}_2 \end{bmatrix} \geq \mathbf{0}. \end{aligned} \quad (194)$$

The resulting approach is referred to as M-EMaC.

A challenging problem when applying M-EMaC is the choice of the parameter m . Intuitively, we need to ensure that the equality holds in (193) for the true signal \mathbf{Z} so that the data information can be appropriately encoded and the frequencies can be correctly recovered from $\mathbf{H}(\mathbf{Z})$. This is guaranteed for a single snapshot once $\mathbf{H}(\mathbf{Z})$ is rank deficient. Unfortunately, a similar argument does not hold in the case of multiple snapshots. In particular, it can be seen from (192) that the rank of $\mathbf{H}(\mathbf{Z})$ also depends on the unknown source signals \mathbf{S} . As an example, in the extreme case when all the sources are coherent, we have that

$$\text{rank}(\mathbf{H}(\mathbf{Z})) \leq \min(K, m, n), \quad (195)$$

which can be much smaller than nL . As a result, if we know that $K < \frac{N}{2}$, then we can set $m = \lceil \frac{N}{2} \rceil$. We leave the parameter tuning in the case of $K \geq \frac{N}{2}$ as an open problem.

6.6 Reweighted Atomic Norm Minimization

In contrast to the EMaC, the atomic norm methods of Subsections 6.3.3 and 6.5.3 can better preserve the signal structure, which is important especially in the noisy case. However, a major disadvantage of the atomic norm is that it suffers from a resolution limit of $\frac{4}{N}$, at least theoretically. To overcome this resolution limit, we present in this subsection the reweighted atomic-norm minimization (RAM) method that was proposed in [84]. The key idea of RAM is to use a smooth surrogate for the atomic ℓ_0 norm, which exploits the sparsity to the greatest extent possible and does not suffer from any resolution limit but is nonconvex and non-smooth, and then optimize the surrogate using a reweighted approach. Interestingly, the resulting reweighted approach is shown to be a reweighted atomic norm with a sound weighting function that gradually enhances sparsity and resolution. While several reweighted approaches have been proposed in the discrete setting (see, e.g., [58, 155–157]), RAM appears to be the first for continuous dictionaries. Since RAM can be applied to single or multiple snapshots as well as to ULA or SLA, we present the result in the most general multiple snapshot SLA case, as in the preceding subsection.

6.6.1 A Smooth Surrogate for $\|\mathbf{Z}\|_{\mathcal{A},0}$

To derive RAM, we first introduce a smooth surrogate for $\|\mathbf{Z}\|_{\mathcal{A},0}$ defined in (180). Note that if the surrogate function is given directly in the continuous frequency domain, then a difficult question is whether and how it can be formulated as a finite-dimensional optimization problem, as $\|\mathbf{Z}\|_{\mathcal{A},0}$ and $\|\mathbf{Z}\|_{\mathcal{A}}$. To circumvent this problem, RAM operates instead in the re-parameterized \mathbf{u} domain. In particular, we have shown that $\|\mathbf{Z}\|_{\mathcal{A},0}$ is equivalent to the following rank minimization problem:

$$\min_{\mathbf{X}, \mathbf{u}} \text{rank}(\mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (196)$$

Inspired by the literature on low rank matrix recovery (see, e.g., [158–160]), the log-det heuristic is adopted as a smooth surrogate for the matrix rank, resulting in the following sparse metric:

$$\mathcal{M}^\epsilon(\mathbf{Z}) := \min_{\mathbf{X}, \mathbf{u}} \log |\mathbf{T}(\mathbf{u}) + \epsilon \mathbf{I}| + \text{Tr}(\mathbf{X}), \text{ subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (197)$$

In (197), ϵ is a tuning parameter that avoids the first term in the objective from being $-\infty$ when $\mathbf{T}(\mathbf{u})$ is rank-deficient, and $\text{Tr}(\mathbf{X})$ is included in the objective to prevent the trivial solution $\mathbf{u} = \mathbf{0}$, as in the SDP in (187) for the atomic norm. Intuitively, as $\epsilon \rightarrow 0$, the log-det heuristic approaches the rank function and $\mathcal{M}^\epsilon(\mathbf{Z})$ would approach $\|\mathbf{Z}\|_{\mathcal{A},0}$. This is indeed true and is formally stated in the following result.

Theorem 6.13 ([84]). *Let $r = \|\mathbf{Z}\|_{\mathcal{A},0}$ and let $\epsilon \rightarrow 0$. Then, we have the following results:*

1. *If $r \leq N - 1$, then*

$$\mathcal{M}^\epsilon(\mathbf{Z}) \sim (r - N) \ln \frac{1}{\epsilon}, \quad (198)$$

i.e., the two quantities above are equivalent infinities. Otherwise, $\mathcal{M}^\epsilon(\mathbf{Z})$ approaches a constant depending only on \mathbf{Z} ;

2. *Let $\hat{\mathbf{u}}_\epsilon$ be the (global) solution \mathbf{u} to the optimization problem in (197). Then, the smallest $N - r$ eigenvalues of $\mathbf{T}(\hat{\mathbf{u}}_\epsilon)$ are either zero or approach zero as fast as ϵ ;*
3. *For any cluster point of $\hat{\mathbf{u}}_\epsilon$ at $\epsilon = 0$, denoted by $\hat{\mathbf{u}}_0$, there exists an atomic decomposition $\mathbf{Z} = \sum_{k=1}^r \mathbf{a}(f_k) \mathbf{s}_k$ such that $\mathbf{T}(\hat{\mathbf{u}}_0) = \sum_{k=1}^r \|\mathbf{s}_k\|_2^2 \mathbf{a}(f_k) \mathbf{a}(f_k)^H$.²*

² \mathbf{u}_0 is called a cluster point of a vector-valued function $\mathbf{u}(x)$ at $x = x_0$ if there exists a sequence $\{x_n\}_{n=1}^{+\infty}$, $\lim_{n \rightarrow +\infty} x_n = x_0$, satisfying $\lim_{n \rightarrow +\infty} \mathbf{u}(x_n) = \mathbf{u}_0$.

Theorem 6.13 shows that the sparse metric $\mathcal{M}^\epsilon(\mathbf{Z})$ approaches $\|\mathbf{Z}\|_{\mathcal{A},0}$ as $\epsilon \rightarrow 0$. Moreover, it characterizes the properties of the optimizer $\hat{\mathbf{u}}_\epsilon$, as $\epsilon \rightarrow 0$, including the convergence speed of the smallest $N - \|\mathbf{Z}\|_{\mathcal{A},0}$ eigenvalues and the limiting form of $\mathbf{T}(\hat{\mathbf{u}}_0)$ via the Vandermonde decomposition. It is worth noting that the term $\ln \frac{1}{\epsilon}$ in (198), which becomes unbounded as $\epsilon \rightarrow 0$, is not problematic in the optimization problem, since the objective function $\mathcal{M}^\epsilon(\mathbf{Z})$ can be re-scaled by $(\ln \frac{1}{\epsilon})^{-1}$ for any $\epsilon > 0$ without altering the optimizer.

In another interesting extreme case when $\epsilon \rightarrow +\infty$, the following result shows that $\mathcal{M}^\epsilon(\mathbf{Z})$ in fact plays the same rule as $\|\mathbf{Z}\|_{\mathcal{A}}$.

Theorem 6.14 ([84]). *Let $\epsilon \rightarrow +\infty$. Then,*

$$\mathcal{M}^\epsilon(\mathbf{Z}) - N \ln \epsilon \sim 2\sqrt{N} \|\mathbf{Z}\|_{\mathcal{A}} \epsilon^{-\frac{1}{2}}, \quad (199)$$

i.e., the two quantities above are equivalent infinitesimals.

As a result, the new sparse metric $\mathcal{M}^\epsilon(\mathbf{Z})$ bridges the atomic norm and the atomic ℓ_0 norm. As ϵ approaches $+\infty$, it approaches the former which is convex and can be globally computed but suffers from a resolution limit. As ϵ approaches 0, it approaches the latter that exploits sparsity to the greatest extent possible and has no resolution limit but cannot be directly computed.

6.6.2 A Locally Convergent Iterative Algorithm

Inserting (197) into (175), we obtain the following optimization problem:

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \log |\mathbf{T}(\mathbf{u}) + \epsilon \mathbf{I}| + \text{Tr}(\mathbf{X}), \\ & \text{subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0} \text{ and } \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F \leq \eta. \end{aligned} \quad (200)$$

This problem is nonconvex since the log-det function is nonconvex. In fact, $\log |\mathbf{T}(\mathbf{u}) + \epsilon \mathbf{I}|$ is a concave function of \mathbf{u} since $\log |\mathbf{R}|$ is a concave function of \mathbf{R} on the positive semidefinite cone [161]. A popular locally convergent approach to the minimization of such a concave + convex function is the majorization-maximization (MM) algorithm (see, e.g., [159]). Let \mathbf{u}_j denote the j th iterate of the optimization variable \mathbf{u} . Then, at the $(j+1)$ th iteration we replace $\ln |\mathbf{T}(\mathbf{u}) + \epsilon \mathbf{I}|$ by its tangent plane at the current value $\mathbf{u} = \mathbf{u}_j$:

$$\begin{aligned} & \ln |\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I}| + \text{Tr} \left[(\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I})^{-1} \mathbf{T}(\mathbf{u} - \mathbf{u}_j) \right] \\ & = \text{Tr} \left[(\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I})^{-1} \mathbf{T}(\mathbf{u}) \right] + c_j, \end{aligned} \quad (201)$$

where c_j is a constant independent of \mathbf{u} . As a result, the optimization problem at the $(j+1)$ th iteration becomes

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \text{Tr} \left[(\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I})^{-1} \mathbf{T}(\mathbf{u}) \right] + \text{Tr}(\mathbf{X}), \\ & \text{subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0} \text{ and } \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F \leq \eta. \end{aligned} \quad (202)$$

Note that the problem in (202) is an SDP that can be globally solved. Since $\log |\mathbf{T}(\mathbf{u}) + \epsilon \mathbf{I}|$ is strictly concave in \mathbf{u} , at each iteration its value decreases by an amount greater than the decrease of its tangent plane. It follows that by iteratively solving (202) the objective function in (200) monotonically decreases and converges to a local minimum.

6.6.3 Interpretation as RAM

We show in this subsection that (202) is actually a weighted atomic norm minimization problem. To do so, let us define a weighted atomic set as (compare with the original atomic set \mathcal{A} defined in (179)):

$$\mathcal{A}^w := \{ \mathbf{a}^w(f, \phi) = w(f) \mathbf{a}(f) \phi : f \in \mathbb{T}, \phi \in \mathbb{C}^{1 \times L}, \|\phi\|_2 = 1 \}, \quad (203)$$

where $w(f) \geq 0$ denotes a weighting function. For $\mathbf{Z} \in \mathbb{C}^{N \times L}$, define its weighted atomic norm as the atomic norm induced by \mathcal{A}^w :

$$\begin{aligned} \|\mathbf{Z}\|_{\mathcal{A}^w} &:= \inf_{c_k, f_k, \phi_k} \left\{ \sum_k c_k : \mathbf{Z} = \sum_k c_k \mathbf{a}^w(f_k, \phi_k), f_k \in \mathbb{T}, \|\phi_k\|_2 = 1, c_k > 0 \right\} \\ &= \inf_{f_k, \mathbf{s}_k} \left\{ \sum_k \frac{\|\mathbf{s}_k\|_2}{w(f_k)} : \mathbf{Z} = \sum_k \mathbf{a}(f_k) \mathbf{s}_k \right\}. \end{aligned} \quad (204)$$

According to the definition above, $w(f)$ specifies the importance of the atom at f : the frequency $f \in \mathbb{T}$ is more likely to be selected if $w(f)$ is larger. The atomic norm is a special case of the weighted atomic norm for a constant weighting function. Similar to the atomic norm, the proposed weighted atomic norm also admits a semidefinite formulation for an appropriate weighting function, which is stated in the following theorem.

Theorem 6.15 ([84]). *Suppose that $w(f) = \frac{1}{\sqrt{\mathbf{a}(f)^H \mathbf{W} \mathbf{a}(f)}} \geq 0$ with $\mathbf{W} \in \mathbb{C}^{N \times N}$. Then,*

$$\begin{aligned} \|\mathbf{Z}\|_{\mathcal{A}^w} &= \min_{\mathbf{X}, \mathbf{u}} \frac{\sqrt{N}}{2} \text{Tr}(\mathbf{W} \mathbf{T}(\mathbf{u})) + \frac{1}{2\sqrt{N}} \text{Tr}(\mathbf{X}), \\ &\text{subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \end{aligned} \quad (205)$$

Let $\mathbf{W}_j = \frac{1}{N} (\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I})^{-1}$ and $w_j(f) = \frac{1}{\sqrt{\mathbf{a}(f)^H \mathbf{W}_j \mathbf{a}(f)}}$. It follows from Theorem 6.15 that the optimization problem in (202) can be exactly written as the following weighted atomic norm minimization problem:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{\mathcal{A}^{w_j}}, \text{ subject to } \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F \leq \eta. \quad (206)$$

As a result, the whole iterative algorithm is referred to as reweighted atomic-norm minimization (RAM). Note that the weighting function is updated based on the latest solution \mathbf{u} . If we let $w_0(f)$ be constant or equivalently, $\mathbf{u}_0 = \mathbf{0}$, such that no preference of the atoms is specified at the first iteration, then the first iteration coincides with ANM. From the second iteration on, the preference is defined by the weighting function $w_j(f)$ given above. Note that $w_j^2(f)$ is nothing but the power spectrum of Capon's beamformer (interpreting $\mathbf{T}(\mathbf{u}_j)$ as the noiseless data covariance and ϵ as the noise variance); also note that similar weighting functions have also appeared in sparse optimization methods in the discrete setting (see, e.g., [58, 63, 157]). The reweighting strategy makes the frequencies around those produced by the current iteration more preferable at the next iteration and thus enhances sparsity. At the same time, the weighting results in resolving finer details of the frequency spectrum in those areas and therefore enhances resolution. In a practical implementation of RAM, we can start with the standard ANM, which corresponds to the case of $\epsilon \rightarrow +\infty$ (by Theorem 6.14), and then gradually decrease ϵ during the iterations.

6.7 Connections between ANM and GLS

We have extended both the atomic norm and the GLS methods from the single to the multiple snapshot case. These two methods were shown in Subsection 6.3.7 to be strongly connected to each other in the single snapshot case, so it is natural to ask whether they are also connected in the multiple snapshot case. We answer this question in this subsection following [82]. In particular, for a small number of snapshots the GLS optimization problem is shown to be equivalent to ANM, whereas for a large number of snapshots it is equivalent to a weighted ANM. Similar results can also be proved for their discrete versions, viz. $\ell_{2,1}$ optimization and SPICE.

6.7.1 The Case of $L < M$

We first consider the ULA case where the GLS optimization problem is given by (165):

$$\min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \widetilde{\mathbf{Y}}^H \\ \widetilde{\mathbf{Y}} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}, \quad (207)$$

where $\widetilde{\mathbf{Y}} = \frac{1}{L} \mathbf{Y} (\mathbf{Y}^H \mathbf{Y})^{\frac{1}{2}}$. By comparing (207) and the SDP formulation of the atomic norm in (187), it can be seen that GLS does nothing but computes $\|\widetilde{\mathbf{Y}}\|_{\mathcal{A}}$ (up to a scaling factor).

A similar argument also holds true in the SLA case where the GLS optimization problem is given by (167):

$$\min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \frac{M}{N} \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \widetilde{\mathbf{Y}}_{\Omega}^H \\ \widetilde{\mathbf{Y}}_{\Omega} & \mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T \end{bmatrix} \geq \mathbf{0}, \quad (208)$$

where $\widetilde{\mathbf{R}}_{\Omega}$ in (167) is replaced here by $\widetilde{\mathbf{Y}}_{\Omega} := \frac{1}{L} \mathbf{Y}_{\Omega} (\mathbf{Y}_{\Omega}^H \mathbf{Y}_{\Omega})^{\frac{1}{2}}$ to reduce the dimensionality. Given $\mathbf{T}(\mathbf{u}) \geq \mathbf{0}$ and applying the identity in (158), we have that

$$\begin{aligned} & \text{Tr}(\widetilde{\mathbf{Y}}_{\Omega}^H [\mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T]^{-1} \widetilde{\mathbf{Y}}_{\Omega}) \\ &= \sum_{t=1}^L \widetilde{\mathbf{Y}}_{\Omega}^H(t) [\mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T]^{-1} \widetilde{\mathbf{Y}}_{\Omega}(t) \\ &= \min_{\mathbf{z}(t)} \sum_{t=1}^L \mathbf{z}^H(t) [\mathbf{T}(\mathbf{u})]^{-1} \mathbf{z}(t), \text{ subject to } \mathbf{z}_{\Omega}(t) = \widetilde{\mathbf{Y}}_{\Omega}(t) \\ &= \min_{\mathbf{Z}} \text{Tr}(\mathbf{Z}^H [\mathbf{T}(\mathbf{u})]^{-1} \mathbf{Z}), \text{ subject to } \mathbf{Z}_{\Omega} = \widetilde{\mathbf{Y}}_{\Omega}. \end{aligned} \quad (209)$$

It follows that

$$\begin{aligned} (208) &\Leftrightarrow \min_{\mathbf{u}} \text{Tr}(\widetilde{\mathbf{Y}}_{\Omega}^H [\mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T]^{-1} \widetilde{\mathbf{Y}}_{\Omega}) + \frac{M}{N} \text{Tr}(\mathbf{T}(\mathbf{u})) \\ &\Leftrightarrow \min_{\mathbf{u}, \mathbf{Z}} \text{Tr}(\mathbf{Z}^H [\mathbf{T}(\mathbf{u})]^{-1} \mathbf{Z}) + \frac{M}{N} \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to } \mathbf{Z}_{\Omega} = \widetilde{\mathbf{Y}}_{\Omega} \\ &\Leftrightarrow \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \frac{M}{N} \text{Tr}(\mathbf{X}) + \frac{M}{N} \text{Tr}(\mathbf{T}(\mathbf{u})), \\ &\text{ subject to } \begin{bmatrix} \mathbf{X} & \sqrt{\frac{N}{M}} \mathbf{Z}^H \\ \sqrt{\frac{N}{M}} \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0} \text{ and } \mathbf{Z}_{\Omega} = \widetilde{\mathbf{Y}}_{\Omega}. \end{aligned} \quad (210)$$

The above SDP is nothing but the following ANM problem (up to a scaling factor):

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{\mathcal{A}}, \text{ subject to } \mathbf{Z}_{\Omega} = \widetilde{\mathbf{Y}}_{\Omega}. \quad (211)$$

We have therefore shown that when $L < M$ the GLS optimization problem is equivalent to certain atomic norm formulations obtained by transforming the observed snapshots. Note that the joint sparsity is preserved in the transformed snapshots in the limiting noiseless case. Therefore, in the absence of noise, by applying the results on the atomic norm, GLS is expected to exactly recover the frequencies under the frequency separation condition. This is true in the ULA case where Theorem 6.11 can be directly applied. However, technically, a similar theoretical guarantee cannot be provided in the SLA case since the assumption on the phases in Theorem 6.12 might not hold true for the transformed source signals.

6.7.2 The Case of $L \geq M$

In this case and for a ULA the GLS optimization problem is given by (see (161)):

$$\min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr}(\widetilde{\mathbf{R}}^{-1} \mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \widetilde{\mathbf{R}}^{\frac{1}{2}} \\ \widetilde{\mathbf{R}}^{\frac{1}{2}} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (212)$$

According to (205), this is nothing but computing the weighted atomic norm $\left\| \widetilde{\mathbf{R}}^{\frac{1}{2}} \right\|_{\mathcal{A}^w}$ (up to a scaling factor), where the weighting function is given by $w(f) = \frac{1}{\sqrt{\mathbf{a}^H(f) \widetilde{\mathbf{R}}^{-1} \mathbf{a}(f)}}$. Note that $w^2(f) = \frac{1}{\mathbf{a}^H(f) \widetilde{\mathbf{R}}^{-1} \mathbf{a}(f)}$ is the power spectrum of the Capon's beamformer.

For an SLA, the GLS problem is given by (see (166)):

$$\min_{\mathbf{X}, \mathbf{u}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{\Gamma}_{\Omega}^T \widetilde{\mathbf{R}}_{\Omega}^{-1} \mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} \mathbf{X} & \widetilde{\mathbf{R}}_{\Omega}^{\frac{1}{2}} \\ \widetilde{\mathbf{R}}_{\Omega}^{\frac{1}{2}} & \mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T \end{bmatrix} \geq \mathbf{0}. \quad (213)$$

By arguments similar to those in the preceding subsection we have that

$$\begin{aligned} (213) &\Leftrightarrow \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{\Gamma}_{\Omega}^T \widetilde{\mathbf{R}}_{\Omega}^{-1} \mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u})), \\ &\text{subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}_{\Omega}^H \\ \mathbf{Z}_{\Omega} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0} \text{ and } \mathbf{Z}_{\Omega} = \widetilde{\mathbf{R}}_{\Omega}^{\frac{1}{2}} \\ &\Leftrightarrow \min_{\mathbf{Z}} \|\mathbf{Z}\|_{\mathcal{A}^w}, \text{ subject to } \mathbf{Z}_{\Omega} = \widetilde{\mathbf{R}}_{\Omega}^{\frac{1}{2}}. \end{aligned} \quad (214)$$

In (214), the weighting function of the weighted atomic norm is given by (up to a scaling factor)

$$w(f) = \frac{1}{\sqrt{\mathbf{a}^H(f) \mathbf{\Gamma}_{\Omega}^T \widetilde{\mathbf{R}}_{\Omega}^{-1} \mathbf{\Gamma}_{\Omega} \mathbf{a}(f)}} = \frac{1}{\sqrt{\mathbf{a}_{\Omega}^H(f) \widetilde{\mathbf{R}}_{\Omega}^{-1} \mathbf{a}_{\Omega}(f)}}. \quad (215)$$

Therefore, here too $w^2(f)$ is the power spectrum of the Capon's beamformer.

We have shown above that for a sufficiently large number of snapshots ($L \geq M$) GLS corresponds to a weighted atomic norm in which the weighting function is given by the square root of the Capon's spectrum. While the standard atomic norm method suffers from a theoretical resolution limit of $\frac{4}{N}$, the above analysis shows that this limit can actually be overcome by using the weighted atomic norm method: indeed GLS is a consistent method and a large-snapshot realization of the MLE. However, it is worth noting that the standard

atomic norm method can be applied to general source signals and its resolution limit is given by a worst case analysis, whereas the statistical properties of GLS are obtained under stronger statistical assumptions on the sources and its performance can degrade if these assumptions are not satisfied. The presented connections between GLS and ANM also imply that GLS is generally robust to source correlations, like ANM, though its power estimates can be biased [82].

6.8 Computational Issues and Solutions

We have presented several gridless sparse methods for DOA/frequency estimation from multiple snapshots. Typically, these methods require computing the solution of an SDP. While several off-the-shelf solvers exist for solving SDP, they generally have high computational complexity. As an example, SDPT3 is an interior-point based method which has the computational complexity of $O(n_1^2 n_2^{2.5})$, where n_1 denotes the number of variables and $n_2 \times n_2$ is the dimension of the PSD matrix of the SDP [162, 163]. To take a look at the computational complexity of the gridless sparse methods, we consider the atomic norm in (187) as an example. Given \mathbf{Z} it can be seen that $n_1 = N + L^2$ and $n_2 = N + L$. So the computational complexity of computing the atomic norm can be rather large, viz. $O\left((N + L^2)^2 (N + L)^{2.5}\right)$. In this subsection we present strategies for accelerating the computation.

6.8.1 Dimensionality Reduction

We show in this subsection that a similar dimensionality reduction technique as introduced in Subsection 4.3.3 can be applied to the atomic norm and the weighted atomic norm methods in the case when the number of snapshots L is large. The technique was firstly proposed in [84] for the gridless setting studied here and it was extended to the discrete case in Subsection 4.3.3. It is also worth noting that a similar dimensionality reduction is not required by GLS since GLS is covariance-based and all the information in the data snapshots \mathbf{Y}_Ω is encoded in the sample covariance matrix $\hat{\mathbf{R}}_\Omega = \frac{1}{L} \mathbf{Y}_\Omega^H \mathbf{Y}_\Omega$ whose dimension does not increase with L . Since it has been shown that GLS and the (weighted) atomic norm are strongly connected, we may naturally wonder if the dimensionality of the atomic norm can be similarly reduced. An affirmative answer is provided in the following result.

Theorem 6.16 ([84]). *Consider the three ANM problems resulting from (175), (176) and (177) which, respectively, are given by:*

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})), \\ & \text{subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0} \text{ and } \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F \leq \eta, \end{aligned} \quad (216)$$

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \lambda' \text{Tr}(\mathbf{X}) + \lambda' \text{Tr}(\mathbf{T}(\mathbf{u})) + \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F^2, \\ & \text{subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}, \end{aligned} \quad (217)$$

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z}} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})), \\ & \text{subject to } \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0} \text{ and } \mathbf{Z}_\Omega = \mathbf{Y}_\Omega. \end{aligned} \quad (218)$$

Let $\widetilde{\mathbf{Y}}_\Omega$ be any matrix satisfying $\widetilde{\mathbf{Y}}_\Omega \widetilde{\mathbf{Y}}_\Omega^H = \mathbf{Y}_\Omega \mathbf{Y}_\Omega^H$, such as the $M \times M$ matrix $(\mathbf{Y}_\Omega \mathbf{Y}_\Omega^H)^{\frac{1}{2}}$. If we replace \mathbf{Y}_Ω by $\widetilde{\mathbf{Y}}_\Omega$ in (216)-(218) and correspondingly change the dimensions of \mathbf{Z} and \mathbf{X} , then the solution \mathbf{u} before and after the replacement is the same. Moreover, if we can find a matrix \mathbf{Q} satisfying $\mathbf{Q}^H \mathbf{Q} = \mathbf{I}$

and $\widetilde{\mathbf{Y}}_\Omega = \mathbf{Y}_\Omega \mathbf{Q}$ and if $(\widehat{\mathbf{X}}, \widehat{\mathbf{Z}}, \widehat{\mathbf{u}})$ is the solution after the replacement, then the solution to the original problems is given by $(\mathbf{Q}\widehat{\mathbf{X}}\mathbf{Q}^H, \widehat{\mathbf{Z}}\mathbf{Q}^H, \widehat{\mathbf{u}})$.

Corollary 6.2 ([84]). *Theorem (6.16) also holds true if the atomic norm is replaced by the weighted atomic norm.*

The dimensionality reduction technique provided by Theorem 6.16 enables us to reduce the number of snapshots from L to M and yet obtain the same solution \mathbf{u} , from which the frequencies and the powers can be retrieved using the Vandermonde decomposition. Therefore, it follows from Theorem 6.16 that for ANM, like for GLS, the information in \mathbf{Y}_Ω is preserved in the sample covariance matrix $\widetilde{\mathbf{R}}_\Omega$. It is interesting to note that the above property even holds true in the presence of coherent sources, while we might expect that DOA estimation from $\widetilde{\mathbf{R}}_\Omega$ could fail in such a case (consider MUSIC as an example).

6.8.2 Alternating Direction Method of Multipliers (ADMM)

A reasonably fast algorithm for SDPs is based on the ADMM [55, 81, 84, 130], which is a first-order algorithm that guarantees global optimality. To derive the ADMM algorithm, we consider (216) as an example. Define

$$\mathcal{S} := \{\mathbf{Z} : \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F \leq \eta\}. \quad (219)$$

Then, (216) can be re-written as:

$$\begin{aligned} \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z} \in \mathcal{S}, \mathbf{Q} \succeq \mathbf{0}} \quad & \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})), \\ \text{subject to } \mathbf{Q} = & \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix}. \end{aligned} \quad (220)$$

We will derive the algorithm following the routine of ADMM by taking $(\mathbf{X}, \mathbf{u}, \mathbf{Z})$ and \mathbf{Q} as the two variables. We introduce $\mathbf{\Lambda}$ as the Lagrangian multiplier and write the augmented Lagrange function for (220) as

$$\begin{aligned} \mathcal{L}(\mathbf{X}, \mathbf{u}, \mathbf{Z}, \mathbf{Q}, \mathbf{\Lambda}) = & \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})) + \text{Tr} \left[\left(\mathbf{Q} - \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \right) \mathbf{\Lambda} \right] \\ & + \frac{\beta}{2} \left\| \mathbf{Q} - \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \right\|_F^2, \end{aligned} \quad (221)$$

where $\beta > 0$ is a penalty parameter set according to [55]. The algorithm is implemented by iteratively updating $(\mathbf{X}, \mathbf{u}, \mathbf{Z})$, \mathbf{Q} and $\mathbf{\Lambda}$ as:

$$(\mathbf{X}, \mathbf{u}, \mathbf{Z}) \leftarrow \arg \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z} \in \mathcal{S}} \mathcal{L}(\mathbf{X}, \mathbf{u}, \mathbf{Z}, \mathbf{Q}, \mathbf{\Lambda}), \quad (222)$$

$$\mathbf{Q} \leftarrow \arg \min_{\mathbf{Q} \succeq \mathbf{0}} \mathcal{L}(\mathbf{X}, \mathbf{u}, \mathbf{Z}, \mathbf{Q}, \mathbf{\Lambda}), \quad (223)$$

$$\mathbf{\Lambda} \leftarrow \mathbf{\Lambda} + \beta \left(\mathbf{Q} - \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \right). \quad (224)$$

Note that \mathcal{L} in (222) is separable and quadratic in \mathbf{X} , \mathbf{u} and \mathbf{Z} . Consequently, these variables can be separately solved for in closed form (note that \mathbf{Z} can be obtained by firstly solving for \mathbf{Z} without the set constraint and then projecting the result onto \mathcal{S}). In (223), \mathbf{Q} can be similarly obtained by firstly solving for \mathbf{Q} without considering the constraint and then projecting the result onto the semidefinite cone, which can be accomplished by forming the eigen-decomposition and setting the negative eigenvalues to zero. The ADMM

algorithm has a per-iteration computational complexity of $O((N + L)^3)$ due to the eigen-decomposition. In the case of $L > M$, this complexity can be reduced to $O((N + M)^3)$ by applying the dimensionality reduction technique presented in the preceding subsection. Although the ADMM may converge slowly to an extremely accurate solution, moderate accuracy is typically sufficient in practical applications [55].

7 Future Research Challenges

In this section we highlight several research challenges that should be investigated in future studies.

- **Improving speed and accuracy:** This is a permanent goal of the research on DOA estimation. As compared to most conventional approaches, in general, the sparse methods may have improved DOA estimation accuracy, especially in difficult scenarios, e.g., the cases with no prior knowledge on the source number, few snapshots and coherent sources. But they are more computationally expensive, which is especially true for the recent gridless sparse methods that typically need to solve an SDP. So efficient SDP solvers should be studied in future research, especially when the array size and the number of snapshots are large. Note that the computation of ℓ_1 optimization has been greatly accelerated during the past decade with the development of compressed sensing.

Furthermore, since the convex sparse methods may suffer from certain resolution limits, it will be of interest to study methods that can directly work with ℓ_0 norms to enhance the resolution. In the case of ULA and SLA, the ℓ_0 optimization corresponds to matrix rank minimization. Therefore, it is possible to apply the recent nonconvex optimization techniques for rank minimization to DOA estimation (see, e.g., [164–169]). Recent progresses in this direction have been made in [170, 171].

- **Automatic model order selection:** Different from the conventional subspace-based methods, the sparse optimization methods usually do not require *a priori* knowledge of the source number (a.k.a. the model order). But this does not mean that the sparse optimization methods are able to accurately estimate the source number. Instead, small spurious sources are usually present in the obtained power spectrum. Therefore, it is of great interest to study how the model order can be automatically estimated within or after the use of the sparse methods. Some results in this direction can be found in [130, 147, 172]. In [130] a parameter refining technique is also introduced once the model order is obtained.
- **Gridless sparse methods for arbitrary array geometry:** The grid-based sparse methods can be applied to any array geometry but they suffer from grid mismatch or other problems. In contrast to this, the gridless sparse methods completely bypass the grid selection problem by utilizing the special Hankel/Toeplitz structure of the sampled data or the data covariance matrix in the case of ULAs and SLAs. For a general array geometry, however, such structures do not exist any more and extension of the gridless sparse methods to general arrays should be studied in future. A recent result in this direction can be found in [173].
- **Gridless sparse parameter estimation and continuous compressed sensing:** Following the line of the previous discussion, it would be of great interest to extend the existing gridless sparse methods for DOA estimation to general parameter estimation problems. Note that a similar data model, as used in DOA estimation, can be formulated for a rather general parameter estimation problem:

$$\mathbf{y} = \sum_{k=1}^K \mathbf{a}(\theta_k) x_k + \mathbf{e}, \quad (225)$$

where \mathbf{y} is the data vector, $\mathbf{a}(\theta)$ is a given function of the continuous parameter θ , x_k are weight coefficients and \mathbf{e} denotes noise. Moreover, to guarantee that the parameters are identifiable as well

as to simplify the model in (225), it is natural to assume that “order” K is small and thus sparsity concept can be introduced as well. But due to the absence of special Hankel/Toeplitz structures, it would be challenging to develop gridless methods for (225). Note that the estimation of θ_k and x_k , $k = 1, \dots, K$ from \mathbf{y} based on the data model in (225) is also referred to as continuous or infinite-dimensional compressed sensing, which extends compressed sensing from the discrete to the continuous setting [127, 154, 174].

8 Conclusions

In this article, we provided an overview of the sparse DOA estimation techniques. Two key differences between sparse representation and DOA estimation were pointed out: 1) discrete system versus continuous parameters and 2) single versus multiple snapshots. Based on how the first difference is dealt with, the sparse methods were classified and discussed in three categories, namely on-grid, off-grid and gridless. The second difference can be tackled by exploiting the temporal redundancy of the snapshots. We explained that while the on-grid and off-grid sparse methods can be applied to arbitrary array geometries, they may suffer from grid mismatch, weak theoretical guarantees etc. These drawbacks can be eliminated by using the gridless sparse methods which, however, can only be applied to ULAs and SLAs. We also highlighted some challenging problems that should be studied in future research. Note that these sparse methods have diverse applications to many fields and the future work also includes performance comparisons of these methods for each specific application. Depending on data qualities and quantities, one or more of these methods may be favored in one application but not another.

References

- [1] J. Capon, High-resolution frequency-wavenumber spectrum analysis, *Proceedings of the IEEE* 57 (8) (1969) 1408–1418.
- [2] V. F. Pisarenko, The retrieval of harmonics from a covariance function, *Geophysical Journal International* 33 (3) (1973) 347–366.
- [3] R. Schmidt, A signal subspace approach to multiple emitter location spectral estimation, Ph.D. thesis, Stanford University (1981).
- [4] R. O. Schmidt, Multiple emitter location and signal parameter estimation, *IEEE Transactions on Antennas and Propagation* 34 (3) (1986) 276–280.
- [5] A. Paulraj, R. Roy, T. Kailath, A subspace rotation approach to signal parameter estimation, *Proceedings of the IEEE* 74 (7) (1986) 1044–1046.
- [6] R. Roy, T. Kailath, ESPRIT-estimation of signal parameters via rotational invariance techniques, *IEEE Transactions on Acoustics, Speech and Signal Processing* 37 (7) (1989) 984–995.
- [7] A. Barabell, Improving the resolution performance of eigenstructure-based direction-finding algorithms, in: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 8, 1983, pp. 336–339.
- [8] H. Krim, M. Viberg, Two decades of array signal processing research: The parametric approach, *IEEE Signal Processing Magazine* 13 (4) (1996) 67–94.

- [9] P. Stoica, R. L. Moses, Spectral analysis of signals, Pearson/Prentice Hall Upper Saddle River, NJ, 2005.
- [10] S. Chen, D. Donoho, M. Saunders, Atomic decomposition by basis pursuit, SIAM Review 43 (1) (2001) 129–159.
- [11] D. L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization, Proceedings of the National Academy of Sciences 100 (5) (2003) 2197–2202.
- [12] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory 52 (2) (2006) 489–509.
- [13] D. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (4) (2006) 1289–1306.
- [14] R. G. Baraniuk, Compressive sensing, IEEE Signal Processing Magazine 24 (4).
- [15] Y. Bresler, A. Macovski, On the number of signals resolvable by a uniform linear array, IEEE Transactions on Acoustics, Speech, and Signal Processing 34 (6) (1986) 1361–1375.
- [16] M. Wax, I. Ziskind, On unique localization of multiple sources by passive sensor arrays, IEEE Transactions on Acoustics, Speech, and Signal Processing 37 (7) (1989) 996–1000.
- [17] A. Nehorai, D. Starer, P. Stoica, Direction-of-arrival estimation in applications with multipath and few snapshots, Circuits, Systems and Signal Processing 10 (3) (1991) 327–342.
- [18] J. Chen, X. Huo, Theoretical results on sparse representations of multiple-measurement vectors, IEEE Transactions on Signal Processing 54 (12) (2006) 4634–4643.
- [19] M. E. Davies, Y. C. Eldar, Rank awareness in joint sparse recovery, IEEE Transactions on Information Theory 58 (2) (2012) 1135–1146.
- [20] Z. Yang, L. Xie, Exact joint sparse frequency recovery via optimization methods, IEEE Transactions on Signal Processing 64 (19) (2016) 5145–5157.
- [21] J. B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear Algebra and Its Applications 18 (2) (1977) 95–138.
- [22] R. G. Baraniuk, E. Candès, R. Nowak, M. Vetterli, Compressive sampling, IEEE Signal Processing Magazine 25 (2) (2008) 12–13.
- [23] R. Chartrand, R. G. Baraniuk, Y. C. Eldar, M. A. Figueiredo, J. Tanner, Introduction to the issue on compressive sensing, IEEE Journal of Selected Topics in Signal Processing 2 (4) (2010) 241–243.
- [24] R. G. Baraniuk, E. Candès, M. Elad, Y. Ma, Applications of sparse representation and compressive sensing [scanning the issue], Proceedings of the IEEE 98 (6) (2010) 906–909.
- [25] J.-L. Starck, J. Fadili, M. Elad, R. D. Nowak, P. Tsakalides, Introduction to the issue on adaptive sparse representation of data and applications in signal and image processing., IEEE Journal of Selected Topics on Signal Processing 5 (5) (2011) 893–895.
- [26] I. F. Gorodnitsky, B. D. Rao, Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm, IEEE Transactions on Signal Processing 45 (3) (1997) 600–616.

- [27] B. D. Rao, K. Kreutz-Delgado, An affine scaling methodology for best basis selection, *IEEE Transactions on signal processing* 47 (1) (1999) 187–200.
- [28] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, K. Kreutz-Delgado, Subset selection in noise based on diversity measure minimization, *IEEE transactions on Signal processing* 51 (3) (2003) 760–770.
- [29] S. Foucart, M. Lai, Sparsest solutions of underdetermined linear systems via ℓ_q -minimization for $0 < q < 1$, *Applied and Computational Harmonic Analysis* 26 (3) (2009) 395–407.
- [30] R. Chartrand, Nonconvex compressed sensing and error correction, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 3, 2007, pp. 889–892.
- [31] R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, *IEEE Signal Processing Letters* 14 (10) (2007) 707–710.
- [32] X. Tan, W. Roberts, J. Li, P. Stoica, Sparse learning via iterative minimization with application to MIMO radar imaging, *IEEE Transactions on Signal Processing* 59 (3) (2011) 1088–1101.
- [33] Y. C. Pati, R. Rezaifar, P. Krishnaprasad, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, in: *1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 40–44.
- [34] D. L. Donoho, Y. Tsaig, I. Drori, J.-L. Starck, Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit, *IEEE Transactions on Information Theory* 58 (2) (2012) 1094–1121.
- [35] J. Tropp, A. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, *IEEE Transactions on Information Theory* 53 (12) (2007) 4655–4666.
- [36] D. Needell, J. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, *Applied and Computational Harmonic Analysis* 26 (3) (2009) 301–321.
- [37] W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, *IEEE Transactions on Information Theory* 55 (5) (2009) 2230–2249.
- [38] M. A. Davenport, M. B. Wakin, Analysis of orthogonal matching pursuit using the restricted isometry property, *IEEE Transactions on Information Theory* 56 (9) (2010) 4395–4401.
- [39] T. Blumensath, M. Davies, Iterative hard thresholding for compressed sensing, *Applied and Computational Harmonic Analysis* 27 (3) (2009) 265–274.
- [40] J. Tropp, S. Wright, Computational methods for sparse solution of linear inverse problems, *Proceedings of the IEEE* 98 (6) (2010) 948–958.
- [41] J. F. Claerbout, F. Muir, Robust modeling with erratic data, *Geophysics* 38 (5) (1973) 826–844.
- [42] E. Candès, Compressive sampling, in: *Proceedings of the International Congress of Mathematicians*, Vol. 3, 2006, pp. 1433–1452.
- [43] E. Candès, The restricted isometry property and its implications for compressed sensing, *Comptes Rendus Mathématique* 346 (9-10) (2008) 589–592.
- [44] S. Foucart, Sparse recovery algorithms: Sufficient conditions in terms of restricted isometry constants, in: *Approximation Theory XIII: San Antonio 2010*, Springer, 2012, pp. 65–77.

- [45] T. Cai, L. Wang, G. Xu, New bounds for restricted isometry constants, *IEEE Transactions on Information Theory* 56 (9) (2010) 4388–4394.
- [46] R. Tibshirani, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society. Series B* 58 (1) (1996) 267–288.
- [47] A. Belloni, V. Chernozhukov, L. Wang, Square-root lasso: pivotal recovery of sparse signals via conic programming, *Biometrika* 98 (4) (2011) 791–806.
- [48] E. Candès, J. Romberg, ℓ_1 -magic: Recovery of sparse signals via convex programming, Available online at <http://users.ece.gatech.edu/~justin/l1magic/downloads/l1magic.pdf>.
- [49] S. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An interior-point method for large-scale ℓ_1 -regularized least squares, *IEEE J. Selected Topics in Signal Processing* 1 (4) (2008) 606–617.
- [50] M. Lustig, D. Donoho, J. Pauly, Sparse MRI: The application of compressed sensing for rapid MR imaging, *Magnetic Resonance in Medicine* 58 (6) (2007) 1182–1195.
- [51] E. T. Hale, W. Yin, Y. Zhang, A fixed-point continuation method for ℓ_1 -regularized minimization with applications to compressed sensing, *CAAM TR07-07*, Rice University 43 (2007) 44.
- [52] Y. Nesterov, Smooth minimization of non-smooth functions, *Mathematical Programming* 103 (1) (2005) 127–152.
- [53] S. Becker, J. Bobin, E. Candès, NESTA: A fast and accurate first-order method for sparse recovery, *SIAM J. Imaging Sciences* 4 (1) (2011) 1–39.
- [54] Z. Yang, C. Zhang, J. Deng, W. Lu, Orthonormal expansion ℓ_1 -minimization algorithms for compressed sensing, *IEEE Transactions on Signal Processing* 59 (12) (2011) 6285–6290.
- [55] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends® in Machine Learning* 3 (1) (2011) 1–122.
- [56] J. Yang, Y. Zhang, Alternating direction algorithms for ℓ_1 -problems in compressive sensing, *SIAM journal on scientific computing* 33 (1) (2011) 250–278.
- [57] D. P. Wipf, B. D. Rao, Sparse Bayesian learning for basis selection, *IEEE Transactions on Signal Processing* 52 (8) (2004) 2153–2164.
- [58] P. Stoica, P. Babu, SPICE and LIKES: Two hyperparameter-free methods for sparse-parameter estimation, *Signal Processing* 92 (7) (2012) 1580–1590.
- [59] M. Tipping, Sparse Bayesian learning and the relevance vector machine, *The Journal of Machine Learning Research* 1 (2001) 211–244.
- [60] S. Ji, Y. Xue, L. Carin, Bayesian compressive sensing, *IEEE Transactions on Signal Processing* 56 (6) (2008) 2346–2356.
- [61] P. Stoica, P. Babu, J. Li, New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data, *IEEE Transactions on Signal Processing* 59 (1) (2011) 35–47.

- [62] P. Stoica, P. Babu, J. Li, SPICE: A sparse covariance-based estimation method for array processing, *IEEE Transactions on Signal Processing* 59 (2) (2011) 629–638.
- [63] P. Stoica, D. Zachariah, J. Li, Weighted SPICE: A unifying approach for hyperparameter-free sparse estimation, *Digital Signal Processing* 33 (2014) 1–12.
- [64] G. McLachlan, T. Krishnan, *The EM algorithm and extensions*, John Wiley & Sons, 1997.
- [65] S. Babacan, R. Molina, A. Katsaggelos, Bayesian compressive sensing using Laplace priors, *IEEE Transactions on Image Processing* 19 (1) (2010) 53–63.
- [66] S. F. Cotter, B. D. Rao, K. Engan, K. Kreutz-Delgado, Sparse solutions to linear inverse problems with multiple measurement vectors, *IEEE Transactions on Signal Processing* 53 (7) (2005) 2477–2488.
- [67] D. Malioutov, M. Cetin, A. Willsky, A sparse signal reconstruction perspective for source localization with sensor arrays, *IEEE Transactions on Signal Processing* 53 (8) (2005) 3010–3022.
- [68] M. Fornasier, H. Rauhut, Recovery algorithms for vector-valued data with joint sparsity constraints, *SIAM Journal on Numerical Analysis* 46 (2) (2008) 577–613.
- [69] M. Mishali, Y. C. Eldar, Reduce and boost: Recovering arbitrary sets of jointly sparse vectors, *IEEE Transactions on Signal Processing* 56 (10) (2008) 4692–4702.
- [70] R. Gribonval, H. Rauhut, K. Schnass, P. Vandergheynst, Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms, *Journal of Fourier Analysis and Applications* 14 (5-6) (2008) 655–687.
- [71] M. Kowalski, Sparse regression using mixed norms, *Applied and Computational Harmonic Analysis* 27 (3) (2009) 303–324.
- [72] S. Ji, D. Dunson, L. Carin, Multitask compressive sensing, *IEEE Transactions on Signal Processing* 57 (1) (2009) 92–106.
- [73] Y. C. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces, *IEEE Transactions on Information Theory* 55 (11) (2009) 5302–5316.
- [74] Y. Eldar, H. Rauhut, Average case analysis of multichannel sparse recovery using convex relaxation, *IEEE Transactions on Information Theory* 56 (1) (2010) 505–519.
- [75] M. Hyder, K. Mahata, Direction-of-arrival estimation using a mixed $\ell_{2,0}$ norm approximation, *IEEE Transactions on Signal Processing* 58 (9) (2010) 4646–4655.
- [76] E. Van Den Berg, M. Friedlander, Theoretical and empirical results for recovery from multiple measurements, *IEEE Transactions on Information Theory* 56 (5) (2010) 2516–2527.
- [77] J. M. Kim, O. K. Lee, J. C. Ye, Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing, *IEEE Transactions on Information Theory* 58 (1) (2012) 278–301.
- [78] K. Lee, Y. Bresler, M. Junge, Subspace methods for joint sparse recovery, *IEEE Transactions on Information Theory* 58 (6) (2012) 3613–3641.
- [79] Y. Chi, L. Scharf, A. Pezeshki, A. Calderbank, Sensitivity to basis mismatch in compressed sensing, *IEEE Transactions on Signal Processing* 59 (5) (2011) 2182–2195.

- [80] D. Chae, P. Sadeghi, R. Kennedy, Effects of basis-mismatch in compressive sampling of continuous sinusoidal signals, in: 2nd IEEE International Conference on Future Computer and Communication (ICFCC), Vol. 2, 2010, pp. 739–743.
- [81] B. N. Bhaskar, G. Tang, B. Recht, Atomic norm denoising with applications to line spectral estimation, *IEEE Transactions on Signal Processing* 61 (23) (2013) 5987–5999.
- [82] Z. Yang, L. Xie, On gridless sparse methods for multi-snapshot DOA estimation, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 3236–3240.
- [83] Y. Li, Y. Chi, Off-the-grid line spectrum denoising and estimation with multiple measurement vectors, *IEEE Transactions on Signal Processing* 64 (5) (2016) 1257–1269.
- [84] Z. Yang, L. Xie, Enhancing sparsity and resolution via reweighted atomic norm minimization, *IEEE Transactions on Signal Processing* 64 (4) (2016) 995–1006.
- [85] B. Ottersten, P. Stoica, R. Roy, Covariance matching estimation techniques for array signal processing applications, *Digital Signal Processing* 8 (3) (1998) 185–210.
- [86] T. Anderson, *Multivariate statistical analysis*, Willey and Sons, New York, NY.
- [87] H. Li, P. Stoica, J. Li, Computationally efficient maximum likelihood estimation of structured covariance matrices, *IEEE Transactions on Signal Processing* 47 (5) (1999) 1314–1323.
- [88] C. Rojas, D. Katselis, H. Hjalmarsson, A note on the SPICE method, *IEEE Transactions on Signal Processing* 61 (18) (2013) 4545–4551.
- [89] P. Babu, P. Stoica, Connection between SPICE and Square-Root LASSO for sparse parameter estimation, *Signal Processing* 95 (2014) 10–14.
- [90] D. P. Wipf, B. D. Rao, An empirical Bayesian strategy for solving the simultaneous sparse approximation problem, *IEEE Transactions on Signal Processing* 55 (7) (2007) 3704–3716.
- [91] Z.-M. Liu, Z.-T. Huang, Y.-Y. Zhou, An efficient maximum likelihood method for direction-of-arrival estimation via sparse Bayesian learning, *IEEE Transactions on Wireless Communications* 11 (10) (2012) 1–11.
- [92] M. Carlin, P. Rocca, G. Oliveri, F. Viani, A. Massa, Directions-of-arrival estimation through Bayesian compressive sensing strategies 61 (7) (2013) 3828–3838.
- [93] P. Stoica, P. Babu, Sparse estimation of spectral lines: Grid selection problems and their solutions, *IEEE Transactions on Signal Processing* 60 (2) (2012) 962–967.
- [94] C. Austin, J. Ash, R. Moses, Dynamic dictionary algorithms for model order and parameter estimation, *IEEE Transactions on Signal Processing* 61 (20) (2013) 5117–5130.
- [95] M. F. Duarte, R. G. Baraniuk, Spectral compressive sensing, *Applied and Computational Harmonic Analysis* 35 (1) (2013) 111–129.
- [96] A. Fannjiang, W. Liao, Coherence pattern-guided compressive sensing with unresolved grids, *SIAM Journal on Imaging Sciences* 5 (1) (2012) 179–202.
- [97] E. J. Candès, C. Fernandez-Granda, Towards a mathematical theory of super-resolution, *Communications on Pure and Applied Mathematics* 67 (6) (2014) 906–956.

- [98] H. Zhu, G. Leus, G. Giannakis, Sparsity-cognizant total least-squares for perturbed compressive sampling, *IEEE Transactions on Signal Processing* 59 (5) (2011) 2002–2016.
- [99] J. Zheng, M. Kaveh, Directions-of-arrival estimation using a sparse spatial spectrum model with uncertainty, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2848–2551.
- [100] Z. Yang, C. Zhang, L. Xie, Robustly stable signal recovery in compressed sensing with structured matrix perturbation, *IEEE Transactions on Signal Processing* 60 (9) (2012) 4658–4671.
- [101] Z. Yang, L. Xie, C. Zhang, Off-grid direction of arrival estimation using sparse Bayesian inference, *IEEE Transactions on Signal Processing* 61 (1) (2013) 38–43.
- [102] Z. Tan, P. Yang, A. Nehorai, Joint sparse recovery method for compressed sensing with structured dictionary mismatch, *IEEE Transactions on Signal Processing* 62 (19) (2014) 4997–5008.
- [103] Y. Zhang, Z. Ye, X. Xu, N. Hu, Off-grid DOA estimation using array covariance matrix and block-sparse Bayesian learning, *Signal Processing* 98 (2014) 197–201.
- [104] M. Lasserre, S. Bidon, O. Besson, F. Le Chevalier, Bayesian sparse Fourier representation of off-grid targets with application to experimental radar data, *Signal Processing* 111 (2015) 261–273.
- [105] W. Si, X. Qu, Z. Qu, Off-grid doa estimation using alternating block coordinate descent in compressed sensing, *Sensors* 15 (9) (2015) 21099–21113.
- [106] T. Chen, H. Wu, L. Guo, L. Liu, A modified rife algorithm for off-grid doa estimation based on sparse representations, *Sensors* 15 (11) (2015) 29721–29733.
- [107] L. Wang, L. Zhao, G. Bi, C. Wan, L. Zhang, H. Zhang, Novel wideband doa estimation based on sparse bayesian learning with dirichlet process priors, *IEEE Transactions on Signal Processing* 64 (2) (2016) 275–289.
- [108] X. Wu, W.-P. Zhu, J. Yan, Direction of arrival estimation for off-grid signals based on sparse bayesian learning, *IEEE Sensors Journal* 16 (7) (2016) 2004–2016.
- [109] Y. Zhao, L. Zhang, Y. Gu, Array covariance matrix-based sparse bayesian learning for off-grid direction-of-arrival estimation, *Electronics Letters* 52 (5) (2016) 401–402.
- [110] G. Han, L. Wan, L. Shu, N. Feng, Two novel DOA estimation approaches for real-time assistant calibration systems in future vehicle industrial, *IEEE Systems Journal*.
- [111] S. Bernhardt, R. Boyer, S. Marcos, P. Larzabal, Compressed sensing with basis mismatch: Performance bounds and sparse-based estimator, *IEEE Transactions on Signal Processing* 64 (13) (2016) 3483–3494.
- [112] Y. Fei, T. Jian-wu, Z. Qing-jie, Off-grid sparse estimator for air velocity in missing-data case, *Journal of Aircraft* (2016) 1–10.
- [113] Q. Shen, W. Cui, W. Liu, S. Wu, Y. D. Zhang, M. G. Amin, Underdetermined wideband doa estimation of off-grid sources employing the difference co-array concept, *Signal Processing* 130 (2017) 299–304.
- [114] J. Yang, G. Liao, J. Li, An efficient off-grid DOA estimation approach for nested array signal processing by using sparse Bayesian learning strategies, *Signal Processing* 128 (2016) 110–122.

- [115] F. Sun, Q. Wu, Y. Sun, G. Ding, P. Lan, An iterative approach for sparse direction-of-arrival estimation in co-prime arrays with off-grid targets, *Digital Signal Processing*.
- [116] D. Shutin, B. H. Fleury, Sparse variational Bayesian SAGE algorithm with application to the estimation of multipath wireless channels, *IEEE Transactions on Signal Processing* 59 (8) (2011) 3609–3623.
- [117] D. Shutin, W. Wang, T. Jost, Incremental sparse Bayesian learning for parameter estimation of superimposed signals, in: *10th International Conference on Sampling Theory and Applications*, 2013.
- [118] L. Hu, Z. Shi, J. Zhou, Q. Fu, Compressed sensing of complex sinusoids: An approach based on dictionary refinement, *IEEE Transactions on Signal Processing* 60 (7) (2012) 3809–3822.
- [119] L. Hu, J. Zhou, Z. Shi, Q. Fu, A fast and accurate reconstruction algorithm for compressed sensing of complex sinusoids, *IEEE Transactions on Signal Processing* 61 (22) (2013) 5744–5754.
- [120] J. Fang, J. Li, Y. Shen, H. Li, S. Li, Super-resolution compressed sensing: An iterative reweighted algorithm for joint parameter learning and sparse signal recovery, *IEEE Signal Processing Letters* 21 (6) (2014) 761–765.
- [121] J. Fang, F. Wang, Y. Shen, H. Li, R. Blum, Super-resolution compressed sensing for line spectral estimation: An iterative reweighted approach, *IEEE Transactions on Signal Processing* 64 (18) (2016) 4649–4662.
- [122] C. Carathéodory, L. Fejér, Über den Zusammenhang der Extremen von harmonischen Funktionen mit ihren Koeffizienten und über den Picard-Landau’schen Satz, *Rendiconti del Circolo Matematico di Palermo* (1884-1940) 32 (1) (1911) 218–239.
- [123] L. Gurvits, H. Barnum, Largest separable balls around the maximally mixed bipartite quantum state, *Physical Review A* 66 (6) (2002) 062311.
- [124] R. A. Horn, C. R. Johnson, *Matrix analysis*, Cambridge University Press, 2012.
- [125] E. J. Candès, C. Fernandez-Granda, Super-resolution from noisy data, *Journal of Fourier Analysis and Applications* 19 (6) (2013) 1229–1254.
- [126] G. Tang, B. N. Bhaskar, P. Shah, B. Recht, Compressive sensing off the grid, in: *50th IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 778–785.
- [127] G. Tang, B. N. Bhaskar, P. Shah, B. Recht, Compressed sensing off the grid, *IEEE Transactions on Information Theory* 59 (11) (2013) 7465–7490.
- [128] G. Tang, B. N. Bhaskar, B. Recht, Near minimax line spectral estimation, *IEEE Transactions on Information Theory* 61 (1) (2015) 499–512.
- [129] Y. Chen, Y. Chi, Robust spectral compressed sensing via structured matrix completion, *IEEE Transactions on Information Theory* 60 (10) (2014) 6576–6601.
- [130] Z. Yang, L. Xie, On gridless sparse methods for line spectral estimation from complete and incomplete data, *IEEE Transactions on Signal Processing* 63 (12) (2015) 3139–3153.
- [131] J.-M. Azais, Y. De Castro, F. Gamboa, Spike detection from inaccurate samplings, *Applied and Computational Harmonic Analysis* 38 (2) (2015) 177–195.

- [132] V. Duval, G. Peyré, Exact support recovery for sparse spikes deconvolution, *Foundations of Computational Mathematics* (2015) 1–41.
- [133] C. Fernandez-Granda, Super-resolution of point sources via convex programming, *Information and Inference*.
- [134] J.-F. Cai, X. Qu, W. Xu, G.-B. Ye, Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction, *Applied and Computational Harmonic Analysis*.
- [135] L. Sun, H. Hong, Y. Li, C. Gu, F. Xi, C. Li, X. Zhu, Noncontact vital sign detection based on stepwise atomic norm minimization, *IEEE Signal Processing Letters* 22 (12) (2015) 2479–2483.
- [136] Z. Yang, L. Xie, C. Zhang, A discretization-free sparse and parametric approach for linear array signal processing, *IEEE Transactions on Signal Processing* 62 (19) (2014) 4959–4973.
- [137] P. Stoica, G. Tang, Z. Yang, D. Zachariah, Gridless compressive-sensing methods for frequency estimation: Points of tangency and links to basics, in: *22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1831–1835.
- [138] B. Sun, H. Feng, Z. Zhang, A new approach for heart rate monitoring using photoplethysmography signals contaminated by motion artifacts, in: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 809–813.
- [139] Y. Zhang, X. Hong, Y. Wang, D. Sun, Gridless SPICE applied to parameter estimation of underwater acoustic frequency hopping signals, in: *2016 IEEE/OES China Ocean Acoustics (COA)*, 2016, pp. 1–6.
- [140] Q. Bao, X. Peng, Z. Wang, Y. Lin, W. Hong, DLSLA 3-d sar imaging based on reweighted gridless sparse recovery method, *IEEE Geoscience and Remote Sensing Letters* 13 (6) (2016) 841–845.
- [141] V. Chandrasekaran, B. Recht, P. A. Parrilo, A. S. Willsky, The convex geometry of linear inverse problems, *Foundations of Computational Mathematics* 12 (6) (2012) 805–849.
- [142] W. Rudin, *Real and complex analysis*, New York: Tata McGraw-Hill Education, 1987.
- [143] B. N. Bhaskar, B. Recht, Atomic norm denoising with applications to line spectral estimation, in: *49th IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2011, pp. 261–268.
- [144] M. Fazel, H. Hindi, S. P. Boyd, A rank minimization heuristic with application to minimum order system approximation, in: *American Control Conference*, Vol. 6, 2001, pp. 4734–4739.
- [145] R. Rochberg, Toeplitz and Hankel operators on the Paley-Wiener space, *Integral Equations and Operator Theory* 10 (2) (1987) 187–235.
- [146] Z. He, A. Cichocki, S. Xie, K. Choi, Detecting the number of clusters in n -way probabilistic clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (11) (2010) 2006–2021.
- [147] Z. Tan, Y. C. Eldar, A. Nehorai, Direction of arrival estimation using co-prime arrays: A super resolution viewpoint, *IEEE Transactions on Signal Processing* 62 (21) (2014) 5565–5576.
- [148] P. Pal, P. Vaidyanathan, A grid-less approach to underdetermined direction of arrival estimation via low rank matrix denoising, *IEEE Signal Processing Letters* 21 (6) (2014) 737–741.

- [149] P. Stoica, T. Soderstrom, On reparametrization of loss functions used in estimation and the invariance principle, *Signal processing* 17 (4) (1989) 383–387.
- [150] D. A. Linebarger, I. H. Sudborough, I. G. Tollis, Difference bases and sparse sensor arrays, *IEEE Transactions on Information Theory* 39 (2) (1993) 716–721.
- [151] P. P. Vaidyanathan, P. Pal, Sparse sensing with co-prime samplers and arrays, *IEEE Transactions on Signal Processing* 59 (2) (2011) 573–586.
- [152] C. Fernandez-Granda, Support detection in super-resolution, *arXiv preprint arXiv:1302.3921*.
- [153] Z. Yang, L. Xie, P. Stoica, Vandermonde decomposition of multilevel Toeplitz matrices with application to multidimensional super-resolution, *IEEE Transactions on Information Theory* 62 (6) (2016) 3685–3701.
- [154] Z. Yang, L. Xie, Continuous compressed sensing with a single or multiple measurement vectors, in: *IEEE Workshop on Statistical Signal Processing (SSP)*, 2014, pp. 308–311.
- [155] M. S. Lobo, M. Fazel, S. Boyd, Portfolio optimization with linear and fixed transaction costs, *Annals of Operations Research* 152 (1) (2007) 341–365.
- [156] E. J. Candes, M. B. Wakin, S. P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *Journal of Fourier Analysis and Applications* 14 (5-6) (2008) 877–905.
- [157] D. Wipf, S. Nagarajan, Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions, *IEEE Journal of Selected Topics in Signal Processing* 4 (2) (2010) 317–329.
- [158] J. David, Algorithms for analysis and design of robust controllers, Ph.D. thesis, Kat. Univ. Leuven (1994).
- [159] M. Fazel, H. Hindi, S. P. Boyd, Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices, in: *American Control Conference*, Vol. 3, 2003, pp. 2156–2162.
- [160] K. Mohan, M. Fazel, Iterative reweighted algorithms for matrix rank minimization, *The Journal of Machine Learning Research* 13 (1) (2012) 3441–3473.
- [161] S. P. Boyd, L. Vandenberghe, *Convex optimization*, Cambridge University Press, London, 2004.
- [162] K.-C. Toh, M. J. Todd, R. H. Tütüncü, SDPT3—a MATLAB software package for semidefinite programming, version 1.3, *Optimization Methods and Software* 11 (1-4) (1999) 545–581.
- [163] L. Vandenberghe, S. Boyd, Semidefinite programming, *SIAM review* 38 (1) (1996) 49–95.
- [164] D. Zachariah, M. Sundin, M. Jansson, S. Chatterjee, Alternating least-squares for low-rank matrix reconstruction, *IEEE Signal Processing Letters* 19 (4) (2012) 231–234.
- [165] S. Bhojanapalli, A. Kyrillidis, S. Sanghavi, Dropping convexity for faster semi-definite optimization, *arXiv preprint*.
- [166] P. Jain, P. Netrapalli, S. Sanghavi, Low-rank matrix completion using alternating minimization, in: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, ACM, 2013, pp. 665–674.

- [167] M. A. Davenport, J. Romberg, An overview of low-rank matrix recovery from incomplete observations, *IEEE Journal of Selected Topics in Signal Processing* 10 (4) (2016) 608–622.
- [168] S. Tu, R. Boczar, M. Soltanolkotabi, B. Recht, Low-rank solutions of linear matrix equations via procrustes flow, *arXiv preprint arXiv:1507.03566*.
- [169] D. Park, A. Kyrillidis, C. Caramanis, S. Sanghavi, Finding low-rank solutions to matrix problems, efficiently and provably, *arXiv preprint arXiv:1606.03168*.
- [170] M. Cho, J.-F. Cai, S. Liu, Y. C. Eldar, W. Xu, Fast alternating projected gradient descent algorithms for recovering spectrally sparse signals, in: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4638–4642.
- [171] J.-F. Cai, T. Wang, K. Wei, Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank Hankel matrix completion, *arXiv preprint arXiv:1606.01567*.
- [172] T. Yardibi, J. Li, P. Stoica, M. Xue, A. B. Baggeroer, Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares, *IEEE Transactions on Aerospace and Electronic Systems* 46 (1) (2010) 425–443.
- [173] K. Mahata, M. M. Hyder, Frequency estimation from arbitrary time samples, *IEEE Transactions on Signal Processing*.
- [174] B. Adcock, A. C. Hansen, Generalized sampling and infinite-dimensional compressed sensing, *Foundations of Computational Mathematics* (2015) 1–61.